Modellierung und Vorhersage von Strukturen biomolekularer Assoziate auf der Basis von statistischen Datenbankanalysen

Vom Fachbereich Chemie
der Technischen Universität Darmstadt

zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigte
Dissertation

vorgelegt von

Dipl.-Ing. Matthias Keil
aus Bensheim

Berichterstatter:	Prof. Dr. J. Brickmann
Mitberichterstatter:	Prof. Dr. J. Gasteiger
	Prof. Dr. H. J. Lindner
Tag der Einreichung:	13. Februar 2002
Tag der mündlichen Prüfung:	29. April 2002

Darmstadt 2002

D 17

Inschrift an der Haustür
meines Elternhauses

Meiner Familie gewidmet

Die vorliegende Arbeit wurde im Fachbereich Chemie, Fachgebiet Physikalische Chemie I, unter Anleitung von Herrn Prof. Dr. J. Brickmann in der Zeit von November 1996 bis Januar 2002 durchgeführt.

Mein Dank an dieser Stelle gilt:

- Herrn Prof. Dr. Jürgen Brickmann für die interessante Themenstellung, die mir gewährte Entfaltungsfreiheit sowie seine stete Unterstützung und Förderung, die es mir u.a. ermöglichte interessante Länder kennenzulernen.

- Herrn Dr. Thomas Exner für seine unermüdliche Diskussionsbereitschaft, viele wissenschaftliche Anregungen, die sorgfältige Durchsicht des Manuskripts und seine motivatorischen Fähigkeiten.

- Herrn Dr. Hans-Jürgen Bär für die gewissenhafte Korrektur dieser Arbeit und die Lösung organisatorischer Probleme.

- Herrn Dipl.-Ing. Thorsten Borosch für sein allzeit offenes Ohr und seine ständige Hilfsbereitschaft.

- Frau Dipl.-Ing. Jamila Saroukh, Herrn Dr. Robert Jäger, Herrn Dr. Dirk Zahn und Herrn Dipl.-Ing. Richard Marhöfer für den regen Austausch, die hervorragende Zusammenarbeit und die unzähligen Capuccino- und Espresso-Pausen.

- Frau Karin Exner für die Beseitigung auch der hartnäckigsten Rechtschreibfehler in dieser Arbeit.

- Allen derzeitigen und ehemaligen Mitgliedern des Arbeitskreises für die freundschaftliche und inspirierende Arbeitsatmosphäre.

- Darüber hinaus danke ich allen weiteren Personen, insbesondere meinen Eltern und meinem Bruder, die zum Gelingen dieser Arbeit beigetragen haben.

Inhaltsverzeichnis

1 Einleitung

Proteine sind in der belebten Natur an einer Vielzahl von biochemischen Prozessen beteiligt. Obwohl sie nur aus einfachen chemischen Bausteinen aufgebaut sind, nehmen sie teilweise hochkomplexe biologische Aufgaben wahr [1]. Die faserförmigen Skleroproteine sind in den Gerüstsubstanzen tierischer Organismen enthalten. So bilden Kollagene den Hauptbestandteil des Stütz- und Bindegewebes (Knochen, Sehnen, Knorpel, etc.). Keratine befinden sich u.a. in Haut, Haaren und Wolle. Die Proteine Actin und Myosin nehmen eine aktive Aufgabe im Muskelgewebe wahr. Sie sind für die Kontraktion der Fibrillen der Muskelzellen verantwortlich. Die Proteine des Immunsystems, die sogenannten Immunglobuline, bilden das Kernstück des biologischen Abwehrsystems höherer Lebewesen. Proteine sind neben diesen Aufgaben auch maßgeblich an der Speicherung, am Transport und der Freisetzung von Stoffwechselenergie beteiligt, regulieren die Genexpression und dienen als Botenstoffe (Hormone) zur Signalübertragung. Als Enzyme sind Proteine an der Katalyse von biochemischen Prozessen beteiligt. Alle Stoffwechselvorgänge, die Biosynthesewege und die Regulation wichtiger physiologischer Prozesse werden durch Enzyme vermittelt. In ihrer Funktion als Biokatalysatoren ermöglichen die Enzyme die Umsetzung des Substrats in das Produkt bei Körpertemperatur und Normaldruck. Das Enzym bindet dabei das Substrat in einer Geometrie, die es auf den Übergangszustand der Reaktion vorbereitet. Durch die Struktur und Anordnung seiner reaktiven Gruppen stabilisiert es den Übergangszustand der chemischen Umsetzung und erniedrigt dadurch die Aktivierungsenergie der Reaktion.

Diese Funktionen der Proteine sind eng mit ihrer Fähigkeit zur Komplexbildung mit anderen Proteinen, DNA-Molekülen oder nichtpeptidischen Liganden verbunden. Die Anlagerung der Proteine an Moleküle ist teilweise hoch selektiv. So lagern sich z.B. Proteine bei der Genexpression nur an bestimmte Basensequenzen des DNA-Stranges an, und Antikörper sind jeweils auf eine Art von Antigenen spezialisiert. Ähnlich selektiv wirken Enzyme. Für fast jede biochemische Reaktion gibt es einen eigenen Biokatalysator, der selektiv ein einziges Molekül oder eine Molekülgruppe umsetzt. Emil Fischer bezeichnete diese Selektivität der Enzyme bildhaft als das Schlüssel-Schloß-Prinzip [2].

Die katalytische Wirkung der Enzyme kann durch stark bindende Moleküle, sogenannte Inhibitoren, verhindert werden. Diese Moleküle blockieren die Enzymfunktion, indem sie sich anstelle des vorgesehenen Substrats an das aktive Zentrum der Enzyme anlagern [3]. Das Verständnis der molekularen Wechselwirkungen, die zur spezifischen Inhibierung von

Enzymen führen, ist von zentraler Bedeutung in der Pharmaforschung, da dadurch die gezielte Entwicklung von Wirkstoffen zur Regulation von physiologischen Prozessen ermöglicht bzw. verbessert werden kann. Die Funktionen der Proteine und Proteinkomplexe sind eng mit ihrer dreidimensionalen Struktur verknüpft. In den letzten Jahren wurden die Strukturen von mehreren tausend Proteinen und Proteinkomplexen aufgeklärt [4,5]. Diese Strukturen können für die Untersuchung der Bindungseigenschaften von Proteinkomplexen verwendet werden. Im Hinblick auf die zur Zeit stattfindenden und teilweise erfolgreich beendeten Entschlüsselungen vollständiger Genome von verschiedenen Organismen wird in der nächsten Zeit mit einer Flut neuer Proteinstrukturen gerechnet, die z.B. aufgrund von Homologiebetrachtungen (*Homology Modeling*) aus der Primärsequenz erzeugt werden können [6].

Bei vielen dieser Proteine sind die Bindungsbereiche und mögliche Bindungspartner nicht bekannt. Methoden zur automatisierten Erkennung von möglichen Bindungsstellen, d.h. aktiven Zentren, in den Proteinstrukturen können bei der gezielten Entwicklung von Wirkstoffen wertvolle Dienste leisten. Vor diesem Hintergrund wurde die in dieser Arbeit vorgestellte computergestützte Methode zur Vorhersage der Bindungsregionen von Proteinkomplexen entwickelt. Dazu sollten zuerst die in der *Protein Data Bank* [4,5] enthaltenen Strukturen mit theoretischen Methoden untersucht und analysiert werden. Dies erforderte die Entwicklung eines automatisierten Verfahrens zur Bearbeitung der mehreren tausend Proteinstrukturen der Proteindatenbank und zur Erzeugung einer ausführlichen Datenbasis. Das Verfahren erlaubt die Untersuchung der Einträge der Proteindatenbank und die Ergänzung von fehlenden Atomen der Proteinstrukturen. Darauf aufbauend können die molekularen Eigenschaften der Proteine (z.B. elektrostatisches Potential, lokale Lipophilie, etc.) berechnet werden. Die Analyse dieser Eigenschaften stützt sich auf das Konzept der molekularen Oberflächen [7-9]. Dabei werden die berechneten Eigenschaften auf die molekularen Oberflächen der Proteine projiziert und die Verteilung der Eigenschaftswerte in den verschiedenen Bereichen der Oberfläche untersucht. Besonderes Gewicht liegt hierbei auf den Bindungsbereichen der Proteinoberfläche zu anderen Molekülen. Dazu sollten umfangreiche statistische Untersuchungen auf der Basis von Datenbankinformationen durchgeführt werden. Aus den Ergebnissen der Analysen lassen sich Rückschlüsse auf die Bindungsmechanismen von Proteinkomplexen ziehen.

Ausgehend von den Ergebnissen dieser Untersuchungen sollte eine Methode zur Vorhersage von Bindungsstellen entwickelt werden. Dazu können neuronale Netzwerke eingesetzt werden. Sie lassen sich auf die Erkennung komplexer Zusammenhänge (z.B.

Mustererkennung) trainieren. In den letzten Jahren wurden immer öfter neuronale Netzwerke zur Lösung von chemischen Fragestellungen herangezogen [10,11]: So verwendet z.B. Gasteiger neuronale Netzwerke zum Vergleich von geometrischen und elektronischen Eigenschaften verschiedener Moleküle [10,12]. Sadowski und Kubinyi entwickelten ein neuronales Netzwerk zur Erkennung von möglichen pharmazeutischen Wirkstoffen anhand von molekularen Deskriptoren [13]. Auch bei der Vorhersage von Sekundärstrukturelementen auf der Basis der Aminosäurensequenz von Proteinen haben sich neuronale Netzwerke bewährt [14-16].

Die vorliegende Arbeit ist wie folgt gegliedert: Nach dieser Einleitung schließt sich ein Abschnitt über die strukturellen Eigenschaften von Proteinkomplexen an. Dort wird auch eine Übersicht über bereits erfolgte Untersuchungen auf diesem Gebiet und die Möglichkeiten zur Vorhersage von Bindungsstellen gegeben. In Kapitel 3 werden die Algorithmen zur Berechnung der molekularen Oberflächen und Eigenschaften der Proteinkomplexe beschrieben. Hierbei werden auch die Methoden, die zur Analyse der Proteineigenschaften neu entwickelt wurden, erläutert. Anschließend wird in Kapitel 4 das automatisierte Verfahren zur Untersuchung von tausenden Proteinstrukturen vorgestellt, und die Analysemethode erläutert. Die Untersuchungsergebnisse werden in Kapitel 5 präsentiert und diskutiert. Dabei wird besonders auf die Unterschiede der Eigenschaften in den Komplexbindungsbereichen und den anderen Oberflächenbereichen der Proteinkomplexe eingegangen. Aufbauend auf diesen Ergebnissen werden in Kapitel 6 Methoden zur Vorhersage von Bindungsbereichen beschrieben. Dabei wird sowohl auf den Aufbau und die Parameter des verwendeten neuronalen Netzwerkes, als auch auf die Vorhersageergebnisse an verschiedenen Proteinkomplexen eingegangen. Alle Ergebnisse der Arbeit werden schließlich in Kapitel 7 zusammengefaßt.

2 Proteinkomplexe

2.1 Strukturelle Eigenschaften von Proteinkomplexen

Die Fähigkeit der Proteine, mit anderen Molekülen Komplexe zu bilden, ist wichtig für ihre biologische Funktion. Proteine können an viele verschiedene Arten von Molekülen binden, dabei sind die Eigenschaften der Bindungsbereiche jeweils an die zu bindenden Moleküle angepaßt. Für die folgenden Untersuchungen werden die vielen verschiedenen Proteinkomplexe grob in drei Gruppen unterteilt:

- Protein-Protein-Komplexe: In dieser Gruppe sind alle Komplexe, in denen mehrere Proteine miteinander binden, zusammengefaßt. Die Proteine können dabei sowohl identische als auch unterschiedliche Aminosäuresequenz besitzen.
- Protein-DNA-Komplexe: Diese Gruppe beinhaltet alle Komplexe, in denen Proteine mit DNA- oder RNA-Molekülen interagieren.
- Protein-Ligand-Komplexe: Komplexe aus Proteinen und beliebigen anderen Molekülen, welche nicht als DNA-/RNA-Moleküle oder Proteine klassifiziert sind.

Im Laufe der letzten Jahre wurden die Strukturen mehrerer tausend Proteine und Proteinkomplexe bestimmt und in der öffentlich zugänglichen *Protein Data Bank* abgelegt [4,5]. Diese dreidimensionalen Strukturdaten können zur Analyse der Eigenschaften von Proteinen und Proteinkomplexen verwendet werden. Durch Vergleich der Bindungsbereiche von Proteinkomplexen mit den nichtbindenden Molekülbereichen können Rückschlüsse auf die Bildungsmechanismen der teilweise hoch spezifischen Komplexe gewonnen werden.

Protein-Protein-Wechselwirkungen bilden die Basis der Quartärstruktur von Proteinen und sind für eine Vielzahl von biochemischen Prozessen wichtig. Dies spiegelt sich auch in der Vielzahl der in der Literatur vorhandenen Untersuchungen von Proteinkomplexstrukturen wider [17-32]. Diese Untersuchungen wurden meist an ausgewählten Proteindimer-, Enzym-Inhibitor- oder Antikörper-Antigen-Komplexen durchgeführt und umfaßten die Analyse einer Vielzahl von Eigenschaften, wie z.B.:

- Aminosäurenzusammensetzung der verschiedenen Proteinbereiche [19,20,22,24-29]
- Hydrophobie der Proteinaußenseite bzw. der Bindungsbereiche [19,20,22,24-29,32-34]

- Elektrostatische Komplementarität im Komplexbindungsbereich [30-32,34]
- Wasserstoffbrückenbindungen zwischen Komplexpartnern [19,20,22,28,31,32]
- Salzbrücken zwischen Komplexpartnern [30-32]
- Größe (molekulare Oberfläche) der Bindungsbereiche [19,20,22,26-28,31,32]
- Anteil einzelner Aminosäuren an der Proteinoberfläche [24,25,35]
- Form und Komplementarität der Bindungsbereiche [22-25,28,34]

Die Ergebnisse der einzelnen Studien unterscheiden sich zwar im Detail, aber alle heben die hohe sterische Komplementarität der Proteine im Komplexbindungsbereich hervor. Weiterhin wird eine Erhöhung von hydrophoben Aminosäuren bzw. Atomgruppen in den Protein-Protein-Bindungsbereichen (Bindungsbereiche von Proteinen zu einem anderen Protein) im Vergleich zu den nichtbindenden Proteinoberflächen beobachtet. Es ist allgemein bekannt, daß der hydrophobe Effekt die treibende Kraft zur Faltung von Proteinen ist [36]. Es wird zudem von mehreren Autoren [18,20,22,27] vermutet, daß hydrophobe Wechselwirkungen auch eine entscheidende Rolle bei der Stabilisierung von Protein-Protein-Komplexen spielen. In den Bindungsbereichen befinden sich laut diesen Untersuchungen jedoch mehr hydrophile Aminosäuren als im Inneren der Proteine. Dies zeige sich auch in dem höherem Anteil von geladenen Aminosäuren. Elcock et al. zeigten [17], daß die Proteinstruktur durch geladene Aminosäuren im Proteininneren destabilisiert wird. Befinden sich die geladenen Aminosäuren im Bindungsbereich von Proteinkomplexen können sie jedoch durch die Bildung von Salzbrücken zwischen den Proteinen zur Komplexstabilisierung beitragen.

Nach Jones et al. [37] sind die Protein-DNA-Bindungsbereiche (Bindungsbereiche von Proteinen zu DNA-Molekülen) im Mittel kleiner als Protein-Protein-Bindungsbereiche. Die Anzahl der Wasserstoffbrückenbindungen ist jedoch höher. Auch die Zahl der hydrophoben Aminosäuren im DNA-Bindungsbereich ist niedriger. Positiv geladene Aminosäuren wie Arginin spielen eine wichtige Rolle bei der Bindung zu den negativ geladenen DNA-Molekülen.

In mehreren Untersuchungen [38-44] wurde deutlich, daß die Protein-Ligand-Bindungsbereiche (Bindungsbereiche von Proteinen zu Ligandmolekülen) in Bezug auf die Anzahl und Verteilung von hydrophilen bzw. hydrophoben Aminosäuren und Wasserstoffbrückenbindungen keine so ausgeprägten Trends zeigen. Jedoch sind Liganden häufig in Taschen oder Spalten der molekularen Oberfläche gebunden, so daß diese Bindungsstellen oft durch einfache geometrische Analysen erkannt werden können.

Zusammengenommen zeigen diese Untersuchungen, daß Proteinkomplexe durch vier Wechselwirkungstypen zwischen den Molekülpartnern charakterisiert werden können:

- Sterische Wechselwirkungen (Van-der-Waals-Kräfte)
- Wasserstoffbrückenbindungen
- Ionische Wechselwirkungen (Salzbrücken)
- Hydrophobe Wechselwirkungen

Die Zahlenwerte der Untersuchungsergebnisse schwanken in den genannten Untersuchungen teilweise stark. Je nach Autor werden z.B. unterschiedliche Größen der Bindungsbereiche, unterschiedliche Anzahl von Wasserstoffbrückenbindungen pro Fläche oder unterschiedliche Einschätzungen der Hydrophobie von Bindungsbereichen angegeben. Sie sind jeweils von den verwendeten Analysemethoden und den zugrundeliegenden Datensätzen der Proteinstrukturen abhängig. Häufig beschränken sich die Arbeiten auf die Untersuchung von einigen wenigen repräsentativen Protein- bzw. Komplexstrukturen (20-100). So wurden von den verschiedenen Autoren unterschiedliche Eigenschaften der Proteinkomplexe ermittelt. Das ist eine direkte Folge der kleinen Zahl von untersuchten Komplexen und unterschiedlichen Ausgangsdaten. Die Ergebnisse an so kleinen Datensätzen dürfen deshalb nicht ohne weiteres auf andere Proteinkomplexe extrapoliert werden. Erst in letzter Zeit wurden einige Untersuchungen mit mehreren hundert Komplexen durchgeführt [26,29,31]. Bis zur Zeit gibt es jedoch keine Untersuchung, die versucht, alle in der *Protein Data Bank* vorhandenen Komplexbindungsbereiche zu analysieren.

2.2 Vorhersage von Proteinbindungsregionen

Bei der Suche nach bzw. der Vorhersage von Bindungsstellen an Proteinoberflächen sind zwei Ausgangssituationen zu beachten, die unterschiedliche Lösungsansätze erfordern. Die konzeptionell einfacher zu lösende Situation liegt vor, wenn die räumlichen Strukturen des Proteins und des Moleküls, welches an das Protein bindet, bekannt sind. Die Verfahren, die zur Vorhersage von Bindungsbereichen unter diesen Vorraussetzungen verwendet werden, sind unter dem Begriff „*Docking*“ zusammengefaßt [3,45-47]. In diesen Verfahren werden zuerst mögliche Anordnungen der Moleküle im Komplex bestimmt und diese dann durch unterschiedliche Funktionen energetisch bewertet. Einige Verfahren verwenden bei der

Suche nach diesen möglichen Komplexstrukturen empirische Ansätze, andere versuchen, durch die globale Optimierung einer Energiefunktion die energetisch beste Komplexstruktur zu bestimmen. Dabei unterscheiden sich die einzelnen Methoden durch die verwendete Energiefunktion und die globale Optimierungsmethode (z.B. Monte-Carlo-Methode, *Simulated Annealing* oder genetische Algorithmen).

Wenn der Komplexpartner nicht bekannt ist, wird die Vorhersage von Bindungsstellen schwieriger, da keine Wechselwirkungsenergien hypothetischer Komplexstrukturen berechnet und zur Evaluation benutzt werden können. Die Analyse von Bindungsbereichen von Proteinkomplexen aus der *Protein Data Bank* zeigt jedoch, wie bereits erläutert, daß die Bindungsstellen von Proteinen von dem Rest der Oberfläche abweichende Eigenschaften besitzen und somit durch die Suche nach diesen speziellen Eigenschaften gefunden werden können.

Für die Entwicklung von neuen Wirkstoffen [3] ist besonders die Erkennung von aktiven Zentren bzw. Bindungsstellen von kleinen Liganden an Proteinoberflächen sehr wichtig. Dementsprechend gibt es in der Literatur eine Menge Arbeiten, die sich mit diesem Problem beschäftigen. Viele dieser Methoden basieren auf der Erkenntnis, daß sich die Bindungsstellen von Liganden bevorzugt in tiefen Taschen oder Spalten der molekularen Proteinoberfläche befinden. Diese Algorithmen beschränken sich deshalb auf rein geometrische Kriterien und suchen die Proteinoberflächen nach charakteristischen Einbuchtungen ab [39,41,48-52]. Dabei werden verschiedenste Vorgehensweisen angewendet. Das Verfahren von Peters et al. zur automatisierten Suche nach Taschen, Spalten und Höhlen in Proteinen [41] basiert auf dem *Alpha-Shape*-Algorithmus von Edelsbrunner [53]. Dieser Algorithmus erzeugt molekulare Oberflächen in verschiedenen Detailstufen. Durch Vergleich der verschiedenen Oberflächen können die Einbuchtungen der molekularen Oberfläche einfach erkannt werden. Das Programm PASS von Brady et al. [39] verteilt in einem mehrstufigen Verfahren Kugeln an der molekularen Oberfläche von Proteinen, wobei Taschen und Spalten durch die Kugeln aufgefüllt, anschließend zu Clustern zusammengefaßt und somit identifiziert werden.

Andere Methoden suchen nach bestimmten Eigenschaften der Proteine an der molekularen Oberfläche (z.B. Polarität, Hydrophobie, Wasserstoffakzeptoren bzw. -donatoren). Das Programm GRID von Goodford [54] ist ein zu diesem Zweck häufig eingesetztes Werkzeug. Es berechnet für funktionelle Gruppen eines potentiellen Liganden günstige Positionen an der molekularen Oberfläche. Dazu wird das zu untersuchende Protein in ein dreidimensionales Gitter positioniert und anschließend Probeteilchen, welche die

physikalischen und chemischen Eigenschaften der funktionellen Gruppen repräsentieren (z.B. Wassermolekül, aromatischer Kohlenstoff, Wasserstoffakzeptor, -donator, etc.), auf den Gitterpunkten plaziert und Wechselwirkungsenergien zu dem Protein berechnet. Durch die Analyse der energetisch günstigsten Positionen können Ligandbindungsstellen an der Proteinaußenseite bestimmt werden. Es gibt viele ähnliche Verfahren, die sich hauptsächlich in der Art der verwendeten Probeteilchen unterscheiden [40,44,55-58].

Neben den bereits vorgestellten Verfahren zur Erkennung möglicher Ligandbindungsstellen an der Proteinoberfläche gibt es auch einige Verfahren zur Vorhersage möglicher Protein-Protein-Bindungsstellen. Der Algorithmus von Young et al. [33] stützt sich auf die Erkenntnis, daß Protein-Protein-Bindungsbereiche hydrophober als der Rest der molekularen Oberfläche sind. Er faßt die hydrophoben Aminosäuren an der Oberfläche zu Clustern zusammen, welche dann der Protein-Protein-Bindungsstelle entsprechen. In 65% der 38 untersuchten Komplexe konnte so die Bindungsstelle korrekt vorhergesagt werden. MacCallum et al. [59] untersuchten Antigen-Antikörper-Komplexe und bestimmten in 26 Komplexen die an der Bindung beteiligten Aminosäuren und benutzten diese Information zur Vorhersage von Antigen-Bindungsstellen. Die Vorhersagemethode von Jones et al. [24,25] beruht auf der Untersuchung von Teilbereichen der Proteinoberflächen. Sie unterteilten die molekulare Oberfläche in überlappende Teilbereiche und bestimmten sechs Parameter, welche die Aminosäurenzusammensetzung, Hydrophobie und Geometrie dieser Oberflächenbereiche charakterisieren. Eine Analyse dieser Parameter an ausgewählten Protein-Protein-Komplexen zeigte, daß sich die Bindungsbereiche von der nichtbindenden Proteinoberfläche deutlich unterscheiden. Jones verwendete die sechs Parameter anschließend zur Vorhersage der Bindungsbereiche von 59 Proteinkomplexen verschiedenen Typs (Oligomere, Enzym-Inhibitor- und Antikörper-Antigen-Komplexe). In 66% dieser Komplexe konnte die korrekte Protein-Protein-Bindungsstelle identifiziert werden.

Einen völlig anderen Weg beschreitet das Verfahren von Lichtarge [60]. Die *Evolutionary Trace Method* basiert auf der Annahme, daß die funktionalen Bereiche und Aminosäuren von Proteinen sich während der Evolution der Proteine nicht wesentlich verändern. D.h. funktional wichtige Aminosäuren eines Proteins sollten in den Aminosäuresequenzen verschiedener Spezies übereinstimmen. Durch eine Analyse der Sequenzen können diese konservierten (invarianten) Aminosäuren bestimmt werden. Konservierte Aminosäuren an der Proteinoberfläche sind sehr häufig an Bindungsstellen bzw. aktiven Zentren beteiligt. Diese Methode wurde in weiteren Arbeiten verfeinert und erweitert [61-63].

3 Molekulare Oberflächen und Eigenschaften

3.1 Molekulare Oberflächen

Chemiker verwenden je nach Problemstellung unterschiedliche Modellkonzepte für die Darstellung von Molekülen. Diese Modelle sind teilweise nicht exakt physikalisch definiert, helfen jedoch bei der Lösung vielfältiger Probleme. Für die Beschreibung von chemischen Reaktionen sind schon einfache zweidimensionale Strichzeichnungen genügend. Benötigt ein Chemiker genaue Informationen über den dreidimensionalen Aufbau von Molekülen, bedient er sich der Draht- oder Kugel-Stab-Modelle aus Molekülbaukästen oder Computerprogrammen. Für viele chemische Probleme ist Wissen über die Größe, das Volumen und die äußere Gestalt der betrachteten Moleküle von entscheidender Bedeutung. Mit Hilfe von molekularen Oberflächen kann die äußere Gestalt von Molekülen dargestellt werden. Moleküle besitzen keine physikalisch definierte Oberfläche, jedoch hat sich das Konzept der molekularen Oberflächen bei dem Studium von Moleküleigenschaften als nützlich erwiesen [3,7,64-67]. Die Untersuchungen in dieser Arbeit stützen sich ebenfalls auf die Analyse der molekularen Oberflächen der Proteinkomplexe. Im Folgenden werden deshalb die verschiedenen Ansätze zur Definition von molekularen Oberflächen [8,9,68-71] und deren Anwendungsmöglichkeiten erläutert.

3.1.1 Isoflächen der Elektronendichteverteilung

Elektronendichteverteilungen von Molekülen können berechnet oder experimentell bestimmt werden. Auf deren Basis lassen sich molekulare Oberflächen durch Berechnung einer oder mehrerer Isoflächen dieser Verteilungen definieren [68]. Ein dabei typisch verwendeter Isowert beträgt 0,002 e^-/a_0^3 (~0,013 $e^-/Å^{-3}$) [72,73]. Für die Berechnung der Elektronendichte gibt es eine Vielzahl quantenmechanischer Methoden. So können je nach gewünschter Genauigkeit und verfügbarer Rechenleistung *ab-initio* oder semiempirische Methoden herangezogen werden. Von verschiedenen Autoren wurden zusätzlich einfach berechenbare Pseudodichtefunktionen zur Generierung molekularer Oberflächen vorgeschlagen [74,75]. Die Verwendung von Isoflächen ist nicht nur auf die Visualisierung von Elektronendichteverteilungen beschränkt. So sind weitere Einsatzgebiete von Isoflächen die Analyse und Visualisierung von Atom- und Molekülorbitalen, elektrostatischen Potentialen, Wechselwirkungsfeldern, und vielem mehr [67].

3.1.2 Van-der-Waals-Oberfläche

Seit langem werden Kalottenmodelle zur Verdeutlichung der Raumerfüllung verwendet. Diese Modelle, oft auch als CPK-Darstellung bezeichnet (nach Corey, Pauling und Koltun), stellen die Atome als Kugeln mit den Van-der-Waals-Radien, die sich aus dem Mindestabstand zweier nichtbindender Atome ergeben, dar. In dieser Arbeit werden die von Bondi zusammengetragenen Radien verwendet [76]. Bei bindenden Atomen durchdringen sich die auf den Atomen zentrierten Kugeln. Die äußere Hülle aller Atomkugeln wird Van-der-Waals-Oberfläche [69] genannt (siehe Abbildung 3.1a).

3.1.3 Zugängliche Oberfläche

Bei der Untersuchung von intermolekularen Wechselwirkungen in Lösungen ist es häufig wichtig, diejenige Oberfläche zu bestimmen, die für ein Lösungsmittelmolekül zugänglich ist. Ausgehend vom Kalottenmodell des Moleküls wird dazu ein Probeteilchen mit vorgegebenem Radius (Standard für das Lösungsmittel Wasser: 1,4 Å [9]) über die Van-der-Waals-Oberfläche gerollt (siehe Abbildung 3.1b). Die Vereinigung aller Mittelpunkte des Probeteilchens wurde erstmals von Lee und Richards beschrieben und *Accessible Surface* genannt [8]. Andererseits verwendet Richards die Oberfläche genau zwischen der Probekugel und dem Kalottenmodell zur Definition der sogenannten *Solvent Accessible Surface* [9]. Sie besteht aus den konvexen Kontaktflächen (*Contact Surface*) der Probekugel mit den Van-der-Waals-Kugeln des Moleküls und den konkaven Flächen (*Reentrant Surface*) dazwischen (siehe Abbildung 3.1). Connolly entwickelte erstmals einen Algorithmus zur Berechnung und Darstellung der *Solvent Accessible Surface* als Punktoberfläche [70]. Für die Visualisierung der Oberfläche und die Berechnung des Oberflächeninhalts und Volumens des Moleküls ist ein Dreiecksnetz, welches diese Punkte verbindet, nützlich. Heiden entwickelte ein Computerprogramm zur nachträglichen Triangulierung von beliebigen Punktoberflächen [7,77]. Dieses kann jedoch bei komplizierten Oberflächen nicht alle Punkte der Oberfläche korrekt miteinander verbinden. Andere Autoren lösen das Problem, indem sie neue von Connolly abweichende Algorithmen implementieren, die bei der Erzeugung der molekularen Oberfläche die Punkte gleichzeitig mit einem Netz von Dreiecken verbinden [78-80]. Neben diesen numerischen Verfahren gibt es aber auch sehr schnelle analytische Methoden zur Berechnung der dem Lösungsmittel zugänglichen molekularen Oberfläche [71,79,81-84].

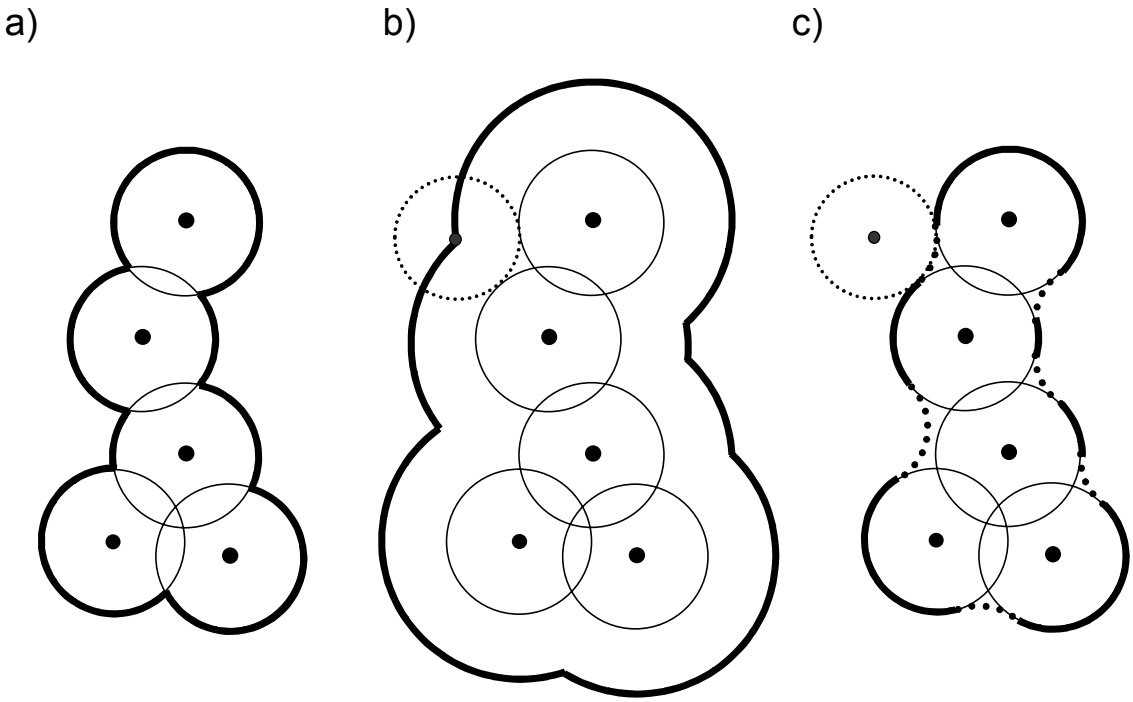


Abbildung 3.1: Molekulare Oberflächendefinitionen:
a) Van-der-Waals-Oberfläche [69]
b) *Accessible Surface* (Lee und Richards [8])
c) *Solvent Accessible Surface* bestehend aus *Reentrant Surface* (···) und *Contact Surface* (—) [70].

3.1.4 Gitterbasierte Berechnung der zugänglichen Oberfläche

In dieser Arbeit wird die *Solvent Accessible Surface* zur Untersuchung der Proteineigenschaften verwendet. Die automatisierte Berechnung einer großen Anzahl von Proteinoberflächen stellt für die bis jetzt vorgestellten Berechnungsmethoden der *Solvent Accessible Surface* ein Problem dar. Das Programm MS von Connolly [70] mit nachgeschalteter Triangulierung der Punktoberfläche mittels TRIADS von Heiden et al. [77] wurde in vielen Untersuchungen von molekularen Wechselwirkungen verwendet, es ist jedoch für die in dieser Arbeit gestellten Aufgaben zu langsam und fehleranfällig. Das Programm MSMS von Sanner et al. [82], welches die *Solvent Accessible Surface* mit einem komplexen, analytischen Algorithmus berechnet, ist sehr schnell. Für einige Moleküle können jedoch mit MSMS keine molekularen Oberflächen erzeugt werden. Deswegen wird hier eine neue numerische Methode basierend auf dem *Marching-Cube*-Algorithmus [85] zur Berechnung molekularer Oberflächen entworfen. Mit Hilfe des *Marching-Cube*-Algorithmus können Isoflächen von dreidimensionalen skalaren Feldern

effizient berechnet werden. Dabei wird die Punktoberfläche und das verbindende Dreiecknetz in einem Schritt erzeugt. Um die dem Lösungsmittel zugängliche molekulare Oberfläche auch mit der Isoflächen-Methode zu erzeugen, ist eine dreidimensionale Funktion nötig, deren Isofläche die gewünschte molekulare Oberfläche ergibt.
Die dreidimensionale Funktion zur Erzeugung der *Accessible Surface* von Lee und Richards [8] ist eine Funktion des Abstandes zum nächstgelegenen Atom und kann wie folgt formuliert werden:

$$D_{vdW}(p_i) = \min_j\left(R_j - d_{ij}\right) \tag{3.1}$$

mit

$D_{vdW}(p_i)$: Abstand des Gitterpunktes p_i zu der Van-der-Waals-Oberfläche des nächstgelegenen Atoms [Å]
d_{ij}: Abstand zwischen Punkt i und Atomzentrum j [Å]
R_j: Van-der-Waals-Radius von Atom j [Å]

Zur Berechnung der Isofläche wird das Molekül zuerst in einem regelmäßigen, kubischen Gitter plaziert. Auf jedem Gitterpunkt wird dann der Abstand $D_{vdW}(p_i)$ zum nächsten Atom gemäß Gleichung 3.1 berechnet. Mit dem Isowert des gewünschten Probeteilchenradius wird anschließend unter Zuhilfenahme des *Marching-Cube*-Algorithmus die *Accessible Surface* als Isofläche erzeugt. Unter Verwendung eines Isowertes von 0 Å erhält man die Van-der-Waals-Oberfläche. Mit der *Accessible Surface* als Zwischenschritt kann auch die gesuchte *Solvent Accessible Surface* berechnet werden (siehe Abbildung 3.2). Hierzu verwendet das Verfahren folgende dreidimensionale Abstandsfunktion:

$$D_{AS}(p_i) = \min_k\left(d_{ik}\right) \tag{3.2}$$

mit

$D_{AS}(p_i)$: Abstand des Gitterpunktes p_i zum nächstgelegenen Punkt der *Accessible Surface* [Å]
d_{ik}: Abstand zwischen Punkt i und Oberflächenpunkt k [Å]

Für jeden Gitterpunkt wird mit Gleichung 3.2 der Abstand zum nächstgelegenen Punkt der zugänglichen Oberfläche berechnet. Mit dem Isowert des Probeteilchenradius werden nun zwei Oberflächen erzeugt. Die eine Oberfläche befindet sich außerhalb der zugänglichen Oberfläche, die andere innerhalb (siehe Abbildung 3.2). Letztere ist eine Näherung der

gesuchten *Solvent Accessible Surface*. Je kleiner die Schrittweite des Gitters gewählt wird, desto ähnlicher wird sie derselben. Tests zeigen, daß schon ab einer Schrittweite des Gitters von 0,5 Å Oberflächen erzeugt werden, die sich von den mit dem Programm MS [70] von Connolly generierten Oberflächen visuell nicht mehr unterscheiden lassen und den Ansprüchen der gestellten Aufgabe in Bezug auf Geschwindigkeit und Automatisierung der Oberflächenerzeugung genügen.

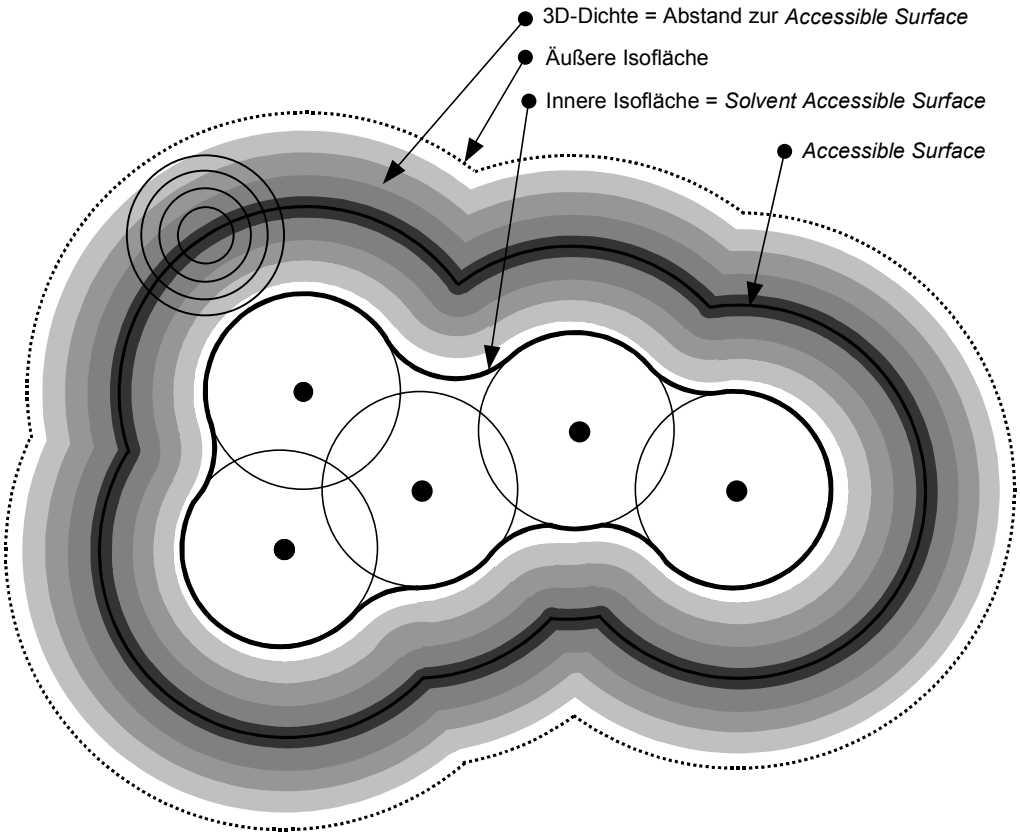


Abbildung 3.2: Gitterbasiertes Berechnungsverfahren für die *Solvent Accessible Surface* auf der Basis einer dreidimensionalen Funktion des Abstandes zur *Accessible Surface*.

3.2 Projektion von Eigenschaften auf die molekularen Oberflächen

Molekulare Oberflächen eignen sich nicht nur zur Analyse der Größe und Form von Molekülen, sondern auch zur Untersuchung von deren Eigenschaften. Dazu werden die molekularen Eigenschaften, wie z.B. das elektrostatische Potential, auf die molekularen Oberflächen projiziert. Für jeden Oberflächenpunkt wird dabei der Wert der gewünschten physikalischen Moleküleigenschaft berechnet und farbkodiert auf der molekularen

Oberfläche dargestellt. Dieses Vorgehen hat sich bei Untersuchungen von Molekül-eigenschaften und Wechselwirkungen bewährt [7,34,86-88]. Im Folgenden werden die auf die Oberfläche projizierten Eigenschaften als „Oberflächeneigenschaften“ bzw. „Eigenschaften der molekularen Oberfläche“ bezeichnet. Von der großen Anzahl von Eigenschaften, die auf die Oberfläche projiziert werden können, werden in dieser Arbeit folgende zur Analyse der Proteinkomplexe herangezogen:

- Elektrostatisches Potential
- Lokale Lipophilie
- Wasserstoffakzeptoren- und Wasserstoffdonatorendichte
- Oberflächenkrümmung
- Tiefe von Taschen in der molekularen Oberfläche
- Molekulare Flexibilität
- Intermolekulare Abstände

3.2.1 Elektrostatisches Potential

Intermolekulare elektrostatische Wechselwirkungen haben entscheidenden Einfluß auf die Bindungsenergie von molekularen Komplexen. Deswegen ist die Untersuchung der elektrostatischen Eigenschaften der Moleküle von elementarer Bedeutung. Das elektrostatische Potential von Molekülen im Vakuum kann mittlerweile für kleine und mittlere Moleküle mit Hilfe von *ab-initio* oder semiempirischen Methoden sehr genau berechnet werden. Für große Moleküle (mehr als wenige hundert Atome) ist der Rechenzeitaufwand mit solchen Methoden immer noch zu hoch. Deshalb muß die Elektrostatik von biochemischen Makromolekülen mit einfacheren Methoden berechnet werden. Die Elektrostatik von Molekülen kann mit der Poisson-Gleichung beschrieben werden:

$$\nabla\varepsilon(\vec{r})\cdot\nabla\varphi(\vec{r}) = -4\pi\rho(\vec{r}) \tag{3.3}$$

mit

$\varphi(\vec{r})$: Elektrostatisches Potential am Ort $\vec{r}$
$\varepsilon(\vec{r})$: Ortsabhängige Dielektrizitätskonstante
$\rho(\vec{r})$: Ladungsdichte

Bei konstanter Dielektrizität ist das Coulomb-Gesetz (Gleichung 3.4) eine Lösung der Poisson-Gleichung (Gleichung 3.3). Es wird oft bei der Berechnung des elektrostatischen Potentials an der Außenseite von kleinen Molekülen im Vakuum benutzt [7]. Dabei können die Partialladungen der Atome über *ab-initio* oder semiempirische Verfahren an kleineren Modellsystemen bestimmt oder Kraftfeldern entnommen werden.

$$\varphi(\vec{r}) = \frac{1}{4\pi\varepsilon_r\varepsilon_0} \cdot \sum_{i=1}^{N} \frac{q_i}{d_i(\vec{r})} \tag{3.4}$$

mit

N: Anzahl der Atome im Molekül

q_i: Partialladung des Atoms *i*

$d_i(\vec{r})$: Abstand des Atoms *i* zum Ort der Potentialberechnung $\vec{r}$

$\varepsilon_r\varepsilon_0$: Dielektrizitätskonstante

Aufgrund der unterschiedlichen Polarisierbarkeit von Wasser und Makromolekülen unterscheiden sich die Dielektrizitätskonstanten von Wasser ($\varepsilon_r = 80$) und Proteinen (2-6) erheblich. Für Untersuchungen in wäßriger Lösung, in der alle biochemischen Reaktionen stattfinden, muß dies bei der Berechnung des elektrostatischen Potentials berücksichtigt und auf die Poisson-Gleichung oder deren Erweiterung, die Poisson-Boltzmann-Gleichung, zurückgegriffen werden [89-91]. Beide Gleichungen modellieren das Lösungsmittel als Kontinuum um das Makromolekül. In der Poisson-Boltzmann-Gleichung wird zusätzlich die Ionenstärke des Lösungsmittels berücksichtigt. Eine analytische Lösung der Poisson- bzw. Poisson-Boltzmann-Gleichung ist für komplexe Systeme wie Makromoleküle nicht möglich. Das elektrostatische Potential wird bei Verwendung dieser Gleichungen deswegen über iterative Algorithmen, vorwiegend das sogenannte Finite-Differenzen-Verfahren, berechnet [89,90,92-94]. Dieses Verfahren ist im Vergleich zur Berechnung mit Hilfe des Coulomb-Gesetzes sehr zeitaufwendig. Mit einer Modifizierung des Coulomb-Ansatzes kann aber die Polarisierbarkeit des Lösungsmittels approximiert werden. In dieser Arbeit wird der *Shifted-Force*-Ansatz von Brooks et al. [95] verwendet. Der Zusatzterm skaliert das elektrostatische Potential, so daß es bis zu einem *Cutoff*-Radius d_0 auf Null absinkt:

$$\varphi(\vec{r}) = \begin{cases} \frac{1}{4\pi\varepsilon} \cdot \sum_{i=1}^{N} \frac{q_i}{d_i} \cdot \left[1 - \left(\frac{d_i}{d_0} \right)^2 \right]^2 & \text{für } d_i \le d_0 \\ 0 & \text{für } d_i > d_0 \end{cases} \tag{3.5}$$

mit

N: Anzahl der Atome im Molekül

q_i: Partialladung des Atoms i

d_i: Abstand des Atoms i zum Ort der Potentialberechnung $\vec{r}$

d_0: *Cutoff*-Radius

Vergleichsrechnungen zeigen qualitative Übereinstimmung der Ergebnisse beider Verfahren. In dieser Arbeit werden nur die Lage und relative Größe von Minima und Maxima des auf die molekulare Oberfläche projizierten elektrostatischen Potentials benötigt. Diese können mit dem modifizierten Coulomb-Ansatz schnell und genügend genau berechnet werden.

3.2.2 Lokale Lipophilie

Kenntnisse von den lokalen lipophilen Eigenschaften eines Moleküls sind ebenso wie dessen Elektrostatik für das Verständnis biochemischer Prozesse wichtig. Verschiedene Ansätze wurden entworfen, um aus der Löslichkeit kleiner organischer Moleküle in polar/unpolaren Zweiphasensystemen die hydrophoben Eigenschaften anhand chemischer Gruppen zu definieren. Als Grundlage zur Quantifizierung der Lipophilie dient hierbei der Verteilungskoeffizient P, der das Verteilungsgleichgewicht einer Substanz zwischen zwei nicht mischbaren Flüssigkeiten beschreibt. Dabei wird meistens das Zweiphasensystem Wasser/n-Oktanol verwendet. Fujita et al. konnten die additiv-konstitutive Natur des Logarithmus des Verteilungskoeffizienten logP nachweisen [96]. Auf Basis dieser Erkenntnis wurden viele Verfahren zur Berechnung von logP-Werten aus Lipophilie-inkrementen entwickelt [97]. Dazu werden die Moleküle in definierte Fragmente zerlegt und anschließend diesen Fragmenten vordefinierte Lipophiliepartialwerte zugeordnet. Die Summe über alle Partialwerte ergibt dann den gesuchten logP-Wert. Manche Verfahren unterteilen die Moleküle in definierte Atomgruppen bzw. funktionelle Gruppen (CLOGP-

Methode), andere weisen jedem Atom des Moleküls einen spezifischen Atomtyp zu (ALOGP-Methode). In beiden Fällen werden die Lipophilieinkremente durch statistische Regressionsverfahren ermittelt.

In dieser Arbeit wird auf dem von Ghose und Crippen entwickelten Verfahren zur Berechnung des logP-Wertes aufgebaut [98,99]. Bei dieser Methode wird die molekulare Struktur durch 110 verschiedene Atomtypen beschrieben. Viswanadhan erweiterte dieses System später auf 120 Atomtypen [100]. In diesen Atomtypen fehlen jedoch geladene Atome, wie sie in den Aminosäuren Arginin, Lysin, Glutaminsäure und Asparaginsäure vorkommen. Deswegen wurden in einer vorhergehenden Arbeit [65] nochmals vier zusätzliche Atomtypen definiert und parametrisiert. Mit diesen Atomparametern wird anschließend der Verteilungskoeffizient logP eines Moleküls über folgende einfache Formel berechnet:

$$\log P = \sum_i f_i \tag{3.6}$$

mit

f_i: Lipophiliepartialwert eines Atoms

Unter der Annahme, daß die atomaren Lipophilieparameter die lokale Lipophilie von Atomen und Atomgruppen repräsentieren, können die Lipophilieinkremente zur Darstellung der lokalen Lipophilie an molekularen Oberflächen verwendet werden. In der Literatur finden sich dazu verschiedene Ansätze: Als ersten Vorschlag führten Audry et al. das *Molecular Lipophilicity Potential* (MLP) ein [101,102]. Dieses ist kein Potential im physikalischen Sinne. Es handelt sich vielmehr um eine abstandsgewichtete Summation der Fragmentwerte.

$$MLP_k = \sum_{i=1}^{N} f_i \cdot \frac{1}{1+\frac{d_{ik}}{d_\Theta}} \qquad \text{wobei} \qquad d_\Theta = 1{,}0\ \text{Å} \tag{3.7}$$

mit

d_{ik}: Abstand von Oberflächenpunkt k zum Molekülfragment i [Å]

f_i: Lipophiliepartialwert des Molekülfragments i

Anstatt der 1/(1+d)-Abstandsabhängigkeit wurden von anderen Autoren weitere Abstandsfunktionen vorgeschlagen. Fauchère et al. [103] verwendeten beispielsweise einen Exponentialansatz:

$$MLP_k = \sum_{i=1}^{N} f_i \cdot e^{-\frac{d_{ik}}{d_\Theta}} \tag{3.8}$$

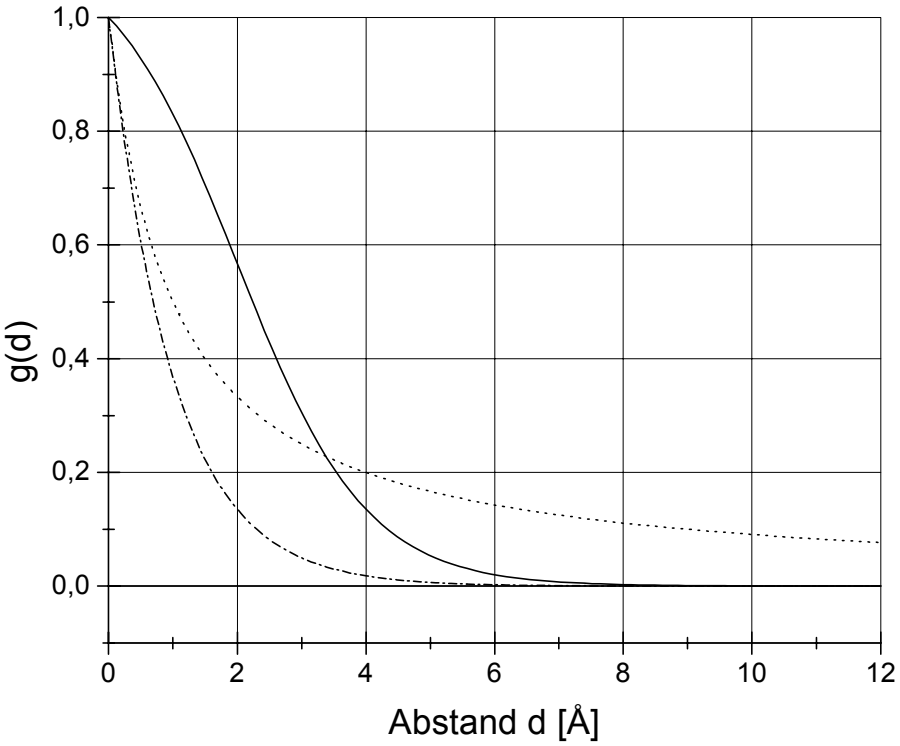
Diese Ansätze (Gleichung 3.7 und 3.8) haben sich bei Untersuchungen von kleinen Molekülen zur Wirkstoffentwicklung bewährt [104,105]. Für Makromoleküle wie Proteine sind diese Ansätze jedoch nicht geeignet, da entweder von der Oberfläche entferntere, im Molekülinneren liegende Atome zu stark berücksichtigt werden (Ansatz von Audry et al. [101]), oder die Abstandsfunktion zu schnell abfällt und somit die lokale Nachbarschaft von Atomen nicht genügend mit einbezogen wird (Ansatz von Fauchère [103]). Heiden und Moeckel [106] entwickelten deshalb eine Abstandsfunktion (Gleichung 3.9), welche die Abschirmung der Kernbereiche großer Moleküle berücksichtigt, indem die Atome in Oberflächennähe stark gewichtet werden, und die Funktion danach schnell abfällt.

$$MLP_{HM} = \frac{\sum_{i=1}^{N} f_i \cdot g_{HM}(d_{ik})}{\sum_{i=1}^{N} g_{HM}(d_{ik})} \quad \text{wobei} \quad g_{HM}(d_{ik}) = \frac{1 + e^{-ab}}{1 + e^{a\left(\frac{d_{ik}}{d_\Theta} - b\right)}} \tag{3.9}$$

mit

a, b: Parameter zur Optimierung des Funktionsverlaufs

Diagramm 3.1 zeigt die Gegenüberstellung der Abstandsfunktionen von Audry et al., Fauchère et al. sowie Heiden et al. [101,103,106]. Die exponentielle Abstandsfunktion fällt zu schnell auf Null ab. Das bedeutet, daß nur die Atome in direkter Nachbarschaft zum betrachteten Oberflächenpunkt in die Berechnung der Lipophilie miteinbezogen werden. Der 1/(1+d)-Term fällt auch sehr schnell ab, nähert sich jedoch erst mit sehr großem Abstand d dem Wert Null, dadurch werden auch Atome in großer Distanz berücksichtigt und beeinflussen die Berechnung der Lipophilie. Gleichung 3.9 fällt im Gegensatz zu den beiden anderen Ansätzen erst ab dem vorgegebenen Abstand stark ab und nähert sich dann rasch dem Nullpunkt.



Diag. 3.1: Abstandsabhängigkeiten der MLP-Berechnungen:
(——): Abstandsabhängigkeit nach Heiden und Moeckel (3.9 mit a = 1,0, b = 2,5) [106].
(- · - ·): Exponentielle Abstandsfunktion nach Fauchère et al. (3.8) [103].
(·····): 1/(1+d) - Abhängigkeit nach Audry et al. (3.7) [101].

3.2.3 Wasserstoffakzeptoren und Wasserstoffdonatoren

Wasserstoffbrücken sind neben sterischen, elektrostatischen und hydrophoben Wechselwirkungen wichtig für viele biochemische Eigenschaften von Proteinen. Sie stabilisieren die Struktur von Proteinen und tragen zur Bildung von Protein-Protein- und Protein-Ligand-Komplexen bei. Wasserstoffbrücken sind Bindungen des Typs X–H···Y mit X als Protonendonator und Y als Protonenakzeptor. X und Y sind elektronegative Atome wie z.B. Stickstoff, Sauerstoff, Halogene, Schwefel und eventuell Phosphor.

3.2.3.1 Wasserstoffbrückenbindungsfähigkeit

Von Heiden wurde ein einfaches Verfahren zur Darstellung der Fähigkeit einzelner Atome zur Bildung von Wasserstoffbrücken auf molekularen Oberflächen vorgestellt [7]. Dazu wird zuerst nach den obigen Kriterien bestimmt, welche Atome eines Moleküls mögliche

Wasserstoffdonatoren bzw. -akzeptoren sind. Anschließend wird für jeden Oberflächenpunkt das nächstliegende Atom ermittelt (kürzester Abstand zur Van-der-Waals-Oberfläche des Atoms). Oberflächenpunkte, die zu Wasserstoffdonatoren oder Wasserstoffakzeptoren gehören, werden dann wie folgt markiert:

−1 = potentieller Wasserstoffakzeptor
+1 = potentieller Wasserstoffdonator
0 = keine Wasserstoffbrückenaktivität

Durch diese Methode erhält man eine Einteilung der molekularen Oberfläche in Wasserstoffakzeptoren- und Wasserstoffdonatorenteiloberflächen. Diese detaillierte Oberflächeninformation ist bei der Untersuchung einzelner Komplexe sehr hilfreich.

3.2.3.2 Wasserstoffakzeptoren- und Wasserstoffdonatorendichte

Bei Kenntnis der Struktur beider Bindungspartner eines Komplexes ist die genaue Lage einzelner Wasserstoffakzeptoren bzw. -donatoren an der Außenseite der Moleküle und deren Anzahl zur Vorhersage möglicher Wasserstoffbrückenbindungen im Komplexbindungsbereich von entscheidender Bedeutung. In der vorliegenden Arbeit wird jedoch die Vorhersage möglicher Bindungsbereiche von Proteinoberflächen ohne Kenntnis des Bindungspartners angestrebt, somit können auch keine Wasserstoffbrückenbindungen im Bindungsbereich auf der Basis der Akzeptor- und Donatorpositionen vorhergesagt werden. Das Wissen über die genaue Lage einzelner Akzeptoren bzw. Donatoren ist daher bei dieser speziellen Problemstellung nur eingeschränkt hilfreich. Informationen über die Anzahl und Verteilung von Wasserstoffakzeptoren und -donatoren in Oberflächenteilbereichen (z.B. Einbuchtungen, Taschen, etc.) liefern jedoch wertvolle Hinweise zur Vorhersage der Bindungsfähigkeit dieser Oberflächenteile.

Basierend auf der oben beschriebenen Einteilung der Oberflächen in Akzeptor- bzw. Donatorteilbereiche kann eine Wasserstoffakzeptorendichte bzw. Wasserstoffdonatorendichte an der molekularen Oberfläche berechnet werden. Zuerst werden dazu, wie oben erläutert, alle Punkte einer molekularen Oberfläche ihren nächsten Atomen zugeordnet und nach ihrer Fähigkeit, Wasserstoffbrücken auszubilden, klassifiziert. Alle Punkte, die einem Atom zugeordnet sind, können zu einer Teiloberfläche (im Folgendem Atom-*Patch* genannt) zusammengefaßt werden. Nun wird für jeden einzelnen Punkt der molekularen Oberfläche die räumliche Umgebung innerhalb eines vorgegebenen Abstandes (*Cutoff*-

Radius) untersucht. Dabei werden die Akzeptoren- bzw. Donatoren-*Patches* innerhalb des gegebenen Radius gezählt und die eingeschlossene molekulare Oberfläche bestimmt. Befindet sich ein Akzeptor bzw. Donator am Rande des untersuchten Oberflächenbereiches, wird nur der Anteil des Atom-*Patch* innerhalb des gewählten Radius gezählt (siehe Abbildung 3.3). Der Quotient aus der ermittelten Anzahl von Wasserstoffakzeptoren bzw. Donatoren und der eingeschlossenen Oberfläche ergibt die gesuchte Dichte:

$$\rho_{acc}(i) = \frac{\sum_j n_{acc}(j)}{a_{gesamt}} \quad \text{bzw.} \quad \rho_{don}(i) = \frac{\sum_j n_{don}(j)}{a_{gesamt}} \tag{3.10}$$

wobei

$$n_{acc/don} = \begin{cases} 1 & \text{Atom} - \textit{Patch} \text{ ist innerhalb des } \textit{Cutoff} - \text{Radius} \\ 0 \le n \le 1 & \text{Atom} - \textit{Patch} \text{ ist nur partiell innerhalb des } \textit{Cutoff} - \text{Radius} \end{cases}$$

mit

$\rho_{acc/don}(i)$: Wasserstoffakzeptorendichte bzw. Wasserstoffdonatorendichte am Punkt i der molekularen Oberfläche [$Å^{-2}$]

$n_{acc/don}(j)$: Anteil des Akzeptors j bzw. Donators j (Atom-*Patch*) innerhalb des *Cutoff*-Radius um den Punkt i

a_{gesamt}: Fläche der im *Cutoff*-Radius um den Punkt i eingeschlossenen molekularen Oberfläche [$Å^2$]

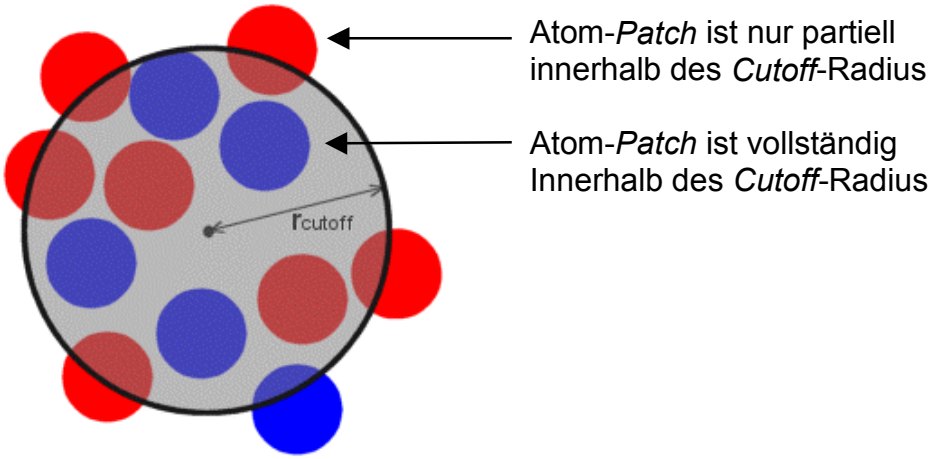


Abbildung 3.3: Berechnung der Wasserstoffakzeptoren- bzw. Wasserstoffdonatorendichte an der Oberfläche von Molekülen (blau: Akzeptoren, rot: Donatoren, grau: *Cutoff*-Radius).

Diese Berechnung kann für Wasserstoffakzeptoren, Wasserstoffdonatoren oder beide gleichzeitig durchgeführt werden. Je nach Wahl des *Cutoff*-Radius erhält man eine mehr oder weniger lokale Information über die Anzahl möglicher Wasserstoffbrücken-bindungsstellen an der Oberfläche von Molekülen.

3.2.3.3 Verteilung von Akzeptoren und Donatoren an der Molekülobefläche

Die Anordnung von Wasserstoffakzeptoren und -donatoren an Molekülobeflächen ist wichtig für die Spezifität von Komplexbindungen. Die Wasserstoffakzeptorendichte bzw. -donatorendichte (Kapitel 3.2.3.2) liefert nur Informationen über die Anzahl von Akzeptoren und Donatoren in Oberflächenbereichen, jedoch keine Informationen, wie diese angeordnet sind, ob z.B. Wasserstoffdonatoren Cluster bilden oder Donatoren und Akzeptoren sehr gleichmäßig immer abwechselnd benachbart vorkommen. Eine Möglichkeit zur Untersuchung der Anordnung ist die Berechnung einer Verteilung der Abstände zwischen Akzeptor/Donator-Paaren an der Molekülaußenseite. Diese radiale Verteilungsfunktion (Paarkorrelationsfunktion [107]) der Akzeptor/Donator-Paare wird wie folgt berechnet: Für jeden Oberflächenteilbereich, der gemäß Kapitel 3.2.3.1 einem Akzeptor oder Donator zugewiesen ist (Wasserstoffbrückenbindungsfähigkeit), wird ein Zentralpunkt auf der molekularen Oberfläche ermittelt. Dazu wird für jeden Punkt des Akzeptor- bzw. Donatoroberflächenbereiches (Atom-*Patch*) der kartesische Abstand zu allen anderen Punkten des Oberflächenbereiches berechnet und aufsummiert. Der Punkt mit der niedrigsten Abstandsumme ist der Zentralpunkt des Oberflächenbereiches. Ausgehend von diesem Punkt i wird nun die Anzahl $n_{i,acc/don}$ der Akzeptoren oder Donatoren (ebenfalls repräsentiert durch jeweils einen Zentralpunkt) in einem Oberflächenring mit dem Radius r und Dicke $\Delta r = 0{,}5$ Å bestimmt (siehe Abbildung 3.4). Die Abstände zwischen den Donator- und Akzeptorzentralpunkten werden dabei nicht mit euklidischer Metrik, sondern auf der Oberfläche entlang der Kanten des Dreiecksnetzes gemessen [64]. Aus der ermittelten Anzahl von Akzeptoren bzw. Donatoren und dem Flächeninhalt des Ringes wird anschließend eine Anzahldichte berechnet und mit der Akzeptor- bzw. Donatordichte $\rho_{acc/don}$ normiert. Die radiale Verteilungsfunktion g(r) ergibt sich aus der Abstandsabhängigkeit dieser Anzahldichte.

$$g(r) = \frac{1}{\rho_{acc/don}} \cdot \frac{n_{i,acc/don}(r)}{A(r)} \qquad (3.11)$$

$$A(r) = F(r + \Delta r) - F(r)$$

mit

g(r): Radiale Verteilungsfunktion

$n_{i,acc/don}(r)$: Anzahl der Akzeptoren oder Donatoren, die sich im Oberflächenring mit dem Radius r und der Dicke Δr um den Zentralpunkt i (Akzeptor oder Donator) befinden.

$\rho_{acc/don}$: Normierung der Verteilungsfunktion: Wasserstoffakzeptoren- oder Wasserstoffdonatorendichte an der Proteinoberfläche

A(r): Flächeninhalt des Ringes mit dem Radius r und der Dicke Δr

F(r): Flächeninhalt des Oberflächenteiles, dessen Oberflächenpunkte einen Abstand zum Zentralpunkt i kleiner als den Radius r besitzen. Die Abstände werden dabei entlang der Oberfläche gemessen.

r: Radius des betrachteten Oberflächenringes

Δr: Dicke des betrachteten Oberflächenringes

Durch den chemischen Aufbau der Aminosäuren sind die Atom-Atom-Abstände von Akzeptoren und Donatoren einer Aminosäure innerhalb bestimmter Grenzen vorgegeben. Daraus können Maxima und Minima in den radialen Verteilungsfunktionen resultieren. Um diesen Effekt zu untersuchen, werden bei der Berechnung zwei verschiedene Methoden angewendet. Einmal werden alle Wasserstoffakzeptoren und -donatoren bei der Berechnung der Verteilung berücksichtigt (Abbildung 3.4 links). Im anderen Fall werden nur die Abstände zwischen Atomen unterschiedlicher Aminosäuren für die Berechnung der Verteilungsfunktion herangezogen (Abbildung 3.4 rechts). Somit werden nur die Akzeptoren-Donatoren-Verteilungsmuster analysiert, die durch die Anordnung der Aminosäuren gegeneinander gebildet werden. In der Abbildung wird die Oberfläche einer Aminosäure durch den grau abgeschatteten Bereich symbolisiert. Die Akzeptor- bzw. Donator-Paare in diesem Bereich werden nicht gezählt. Ebenso wird die Ringfläche dieser Aminosäure in die Berechnung der radialen Verteilungsfunktion nicht einbezogen. Insgesamt werden vier radiale Verteilungsfunktionen bestimmt: Akzeptor-Akzeptor-, Akzeptor-Donator-, Donator-Akzeptor- und Donator-Donator-Abstände.

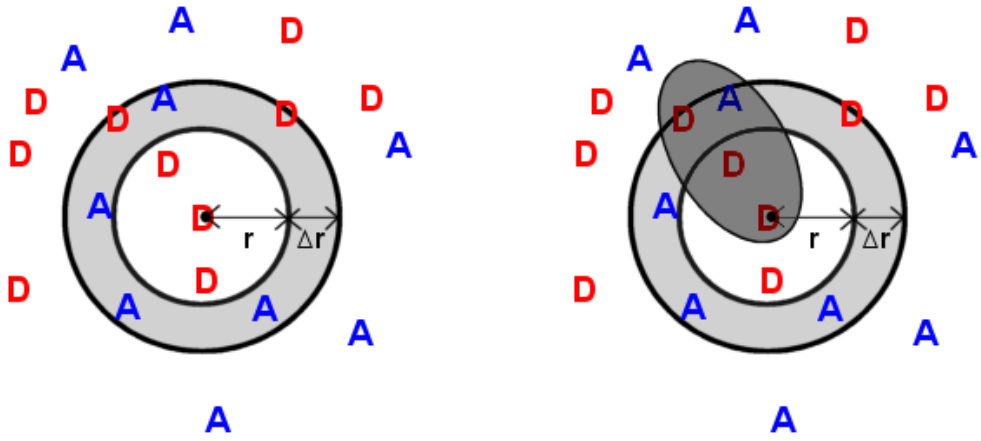


Abb. 3.4: Verteilung von Wasserstoffakzeptoren und Wasserstoffdonatoren an der molekularen Oberfläche:
links) Berechnung der radialen Verteilungsfunktion an der molekularen Oberfläche unter Berücksichtigung aller Akzeptoren und Donatoren.
rechts) Berechnung ohne Berücksichtigung der Akzeptoren bzw. Donatoren in der gleichen Aminosäure. Der abgeschattete Bereich symbolisiert den Oberflächenbereich einer Aminosäure. Nur die Akzeptoren/Donatoren bzw. die Ringfläche im hellgrauen Bereich des Ringes werden berücksichtigt.

3.2.4 Oberflächenkrümmung

Für die Ausbildung von Wasserstoffbrücken, elektrostatischen und lipophilen Wechselwirkungen in einer starken Komplexbindung muß sterische Komplementarität vorliegen, d.h. die Form der molekularen Oberfläche beider Moleküle muß zueinander passen. Damit wird die Kontaktfläche der beiden Moleküle maximiert, und es können sich die oben genannten Wechselwirkungen zwischen den Atomen der Komplexpartnern ausbilden und so zur Stabilisierung des Komplexes beitragen. Dieses Prinzip wurde erstmals von Emil Fischer formuliert und als Schlüssel-Schloß-Prinzip bezeichnet [2].

Die lokale Topographie von Oberflächen kann über die kanonischen Krümmungen quantitativ beschrieben werden [108]. Die kanonischen Krümmungen sind als Eigenwerte der lokalen Hesse-Matrix definiert und beschreiben die direkte Umgebung eines Oberflächenpunktes in atomarer Auflösung. Für die Beschreibung der Wechselwirkungen von biochemischen Komplexen ist jedoch eine Beschreibung der Topographie größerer zusammenhängender Molekülbereiche, wie z.B. einer Komplexbindungsstelle erwünscht. Von Zachmann [109] wurde die Definition der kanonischen Krümmungen auf Oberflächenbereiche vordefinierter Größe erweitert. Zur Berechnung dieser globalen Krümmungen werden zu jedem Oberflächenpunkt alle Punkte innerhalb eines

vorgegebenen Maximalabstandes (*Cutoff*-Radius) bestimmt und an diese mit Hilfe eines numerischen Verfahrens ein Paraboloid angenähert. Die Hauptkrümmungen des Paraboloids definieren die globalen Krümmungen der Oberfläche an jedem Oberflächenpunkt. In konvexen Bereichen sind die Werte der Krümmungen negativ und in konkaven Bereichen positiv definiert. Dadurch sind konvexe Regionen der molekularen Oberfläche durch zwei negative, konkave Regionen durch zwei positive globale Krümmungen festgelegt. Eine negative und eine positive Krümmung beschreiben eine sattelförmige Oberfläche.

Basierend auf den beiden globalen Krümmungen wurde von Heiden [7] der *Surface Topography Index* (STI) definiert. Er beschreibt die Oberflächenform anhand fünf Basisformen und unterscheidet zwischen gleichmäßig konkaven (Loch, STI=0), langgestreckt konkaven (Spalt, STI=1), sattelförmigen (Sattel, STI=2), langgestreckt konvexen (Grat, STI=3) bzw. gleichmäßig konvexen Oberflächenbereichen (Pfropf, STI=4). Der STI wird an jedem Oberflächenpunkt aus den beiden globalen Krümmungen k_1 und k_2 wie folgt berechnet:

$$\text{STI} = \frac{k_1 - k_2}{k_1} \qquad \text{wenn } k_1 > 0 \text{ und } k_2 > 0 \text{ oder wenn } k_1 > 0,\ k_2 \le 0 \text{ und } |k_1| > |k_2|$$

$$\text{STI} = \frac{k_1 + 3 \cdot k_2}{k_2} \qquad \text{wenn } k_1 > 0 \text{ und } k_2 \le 0,\ |k_1| \le |k_2| \text{ oder wenn } k_1 \le 0 \text{ und } k_2 < 0$$

Der Sonderfall einer ebenen Oberflächenregion, wenn also beide globalen Krümmungen gleich Null sind, wird mit einem STI-Wert von -1 berücksichtigt.

$$\text{STI} = -1 \qquad \text{wenn } k_1 = k_2 = 0$$

3.2.5 Tiefeninformation

Häufig befinden sich Ligandbindungsstellen und aktive Zentren von Proteinen in Taschen und Spalten der molekularen Oberfläche. Insbesondere kleinere Liganden sind so gebunden. Es existieren viele Ansätze zur automatischen Bestimmung von solchen Bindungsregionen an der Oberfläche von Proteinen. Levitt and Banaszak [48] entwickelten ein interaktives Programm zur Identifizierung, Darstellung und Manipulation von Taschen in bekannten Proteinstrukturen. Dabei wird eine Probekugel entlang der x-, y- und z-Achse

verschoben und die Positionen, an denen die Kugel in beiden Richtungen der Achse von Proteinatomen umgeben ist, jedoch kein Atomzentrum berührt, als Taschenbereich definiert. Ähnliche Ansätze [50,51] benutzen Probekugeln, deren Größe verändert werden, bis sie die nächstgelegenen Atome berühren. Die Taschen werden dann über die Vereinigungsmenge der überlappenden Probekugeln definiert. Die Methode von Delaney [49] benutzt einen Mustererkennungsalgorithmus basierend auf zellulären Logik-operationen zur Unterscheidung zwischen konvexen und konkaven Regionen von Proteinen. Von Exner et al. [52] wurde ein gitterbasierter Algorithmus vorgeschlagen. Gitterpunkte werden als Punkte innerhalb einer Proteintasche definiert, sofern sie in mindestens zwei Richtungen entlang der Achsen des kartesischen Gitters von Punkten innerhalb des Proteins umgeben sind. Diese Punkte werden dann zu Clustern, die Taschen bzw. Spalten bilden, zusammengefaßt.

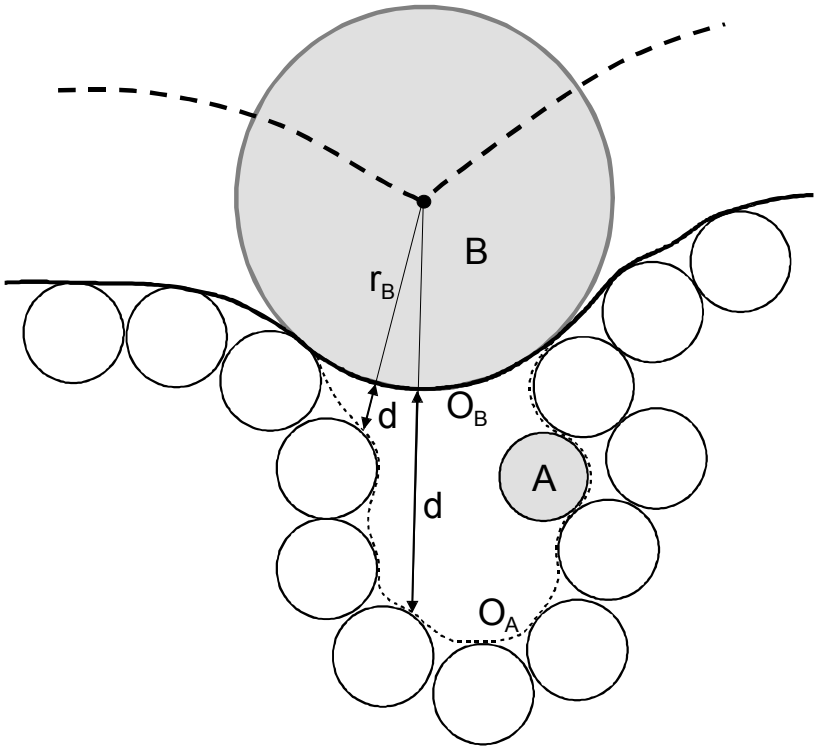


Abbildung 3.5: Berechnung der Tiefe von Taschen in der molekularen Oberfläche.

In dieser Arbeit wird eine neue Methode zur Berechnung der Tiefe von Spalten und Taschen der Proteinoberflächen entwickelt und verwendet (siehe Abbildung 3.5). Für das untersuchte Protein werden dazu zwei molekulare Oberflächen mit unterschiedlichen Radien r_A bzw. r_B der Probekugel berechnet. Für die erste molekulare Oberfläche O_A wird der Standardradius r_A = 1,4 Å und für die zusätzliche molekulare Oberfläche O_B eine

größere Probekugel B ($r_B = 6,0$) verwendet. So werden Taschen und Spalten, in welche die größere Probekugel nicht hineinpaßt, abgedeckt. Nun wird für jeden Punkt der Oberfläche O_A der Abstand d zur zweiten Oberfläche O_B berechnet. Dazu wird der Abstand jedes Oberflächenpunktes zum Mittelpunkt der am nächsten gelegenen Probekugel B bestimmt und der Radius r_B subtrahiert. Dadurch ist gewährleistet, daß der Vektor der Abstandsmessung senkrecht auf der Oberfläche O_B steht. Diese Distanz ist ein Maß für die Tiefe der betreffenden Tasche oder Spalte in der molekularen Oberfläche. Ab einem Abstand d von 2,5 Å wird dieser Oberflächenpunkt als ein Punkt innerhalb einer Oberflächenspalte betrachtet. Durch Entfernen der restlichen Punkte (Abstand kleiner 2,5 Å) werden Oberflächenteilstücke, die mögliche Ligandbindungsstellen darstellen, isoliert.
Mit dieser Methode werden auch Kanäle und innere Oberflächen (keine Verbindung zur Außenseite der Proteine) erkannt, jedoch nicht entsprechend klassifiziert. Dies kann recht einfach durch Betrachtung der Normalenvektoren der Oberflächenpunkte erzielt werden. Der Betrag des gemittelten Normalenvektors von Spalten und Taschen ist sehr viel höher als von zylindrischen Kanälen und Hohlräumen innerhalb der Proteine. Als Grenzwert für die Unterscheidung zwischen Taschen und Hohlräumen wird ein Betrag von 0,2 verwendet. Hat der gemittelte Normalenvektor einen Betrag kleiner 0,2, handelt es sich um einen Kanal oder abgeschlossenen Hohlraum innerhalb des Proteins.

3.2.6 Molekulare Flexibilität

Moleküle sind keine starren Objekte. Insbesondere an der Oberfläche der Moleküle können einzelne Atomgruppen sehr beweglich sein. Zachmann entwickelte zwei Verfahren zur Darstellung der molekularen Flexibilitäten auf molekularen Oberflächen [110,111]. Das erste Verfahren beruht auf einer statistischen Analyse der molekularen Oberflächen eines Proteins, dessen Flexibilität mittels einer molekulardynamischen Simulation (MD-Simulation) untersucht wurde. Die zweite Methode ist eine abstandsgewichtete Darstellung der gemittelten Fluktuationen der Atompositionen auf der molekularen Oberfläche. Als Abstandsfunktion wird dabei wiederum Gleichung 3.9 verwendet. Diese Fluktuationen können durch MD-Simulationen bestimmt werden oder sind bei den experimentellen Daten einer Kristallstrukturbestimmung als Debye-Waller-Temperaturfaktoren (B-Faktoren) angegeben. Die aus MD-Simulationen gewonnenen atomaren Beweglichkeiten sind über Gleichung 3.11 mit den experimentell ermittelten B-Faktoren verknüpft. In dieser Arbeit wird nur das zweite auf den Temperaturfaktoren basierende Verfahren verwendet, da MD-Simulationen von Makromolekülen sehr zeitaufwendig sind und somit die entsprechende

Behandlung der großen Anzahl von Proteinstrukturen im Rahmen dieser Arbeit nicht durchführbar ist.

$$\left\langle \Delta r_i^2 \right\rangle = \frac{3B_i}{8\pi^2} \tag{3.12}$$

mit

$\left\langle \Delta r_i^2 \right\rangle$: quadratische gemittelte Fluktuation des Atoms i aus Molekulardynamik-Simulationen [$Å^2$]

B_i: Debye-Waller-Temperaturfaktor des Atoms i

3.2.7 Intermolekulare Abstände

Die Berechnung von Abständen der molekularen Oberfläche zu anderen Molekülen oder Atomgruppen ist hilfreich für die automatische Bestimmung der Bindungsregionen von Komplexen. Wenn der berechnete Abstand einen gegebenen Maximalabstand (1,5 Å) nicht übersteigt, wird der betreffende Oberflächenbereich als Bindungsregion markiert. Diese Teiloberflächen können dann vermessen und untersucht werden. Form, Größe und physikochemische Eigenschaften der Bindungsflächen sind wichtige Parameter bei den folgenden Untersuchungen. Durch Abstandsberechnungen zwischen den beteiligten Komplexpartnern wird zusätzlich bestimmt, welche Atomgruppen bzw. Aminosäuren sich bevorzugt in Bindungsregionen von Proteinkomplexen befinden.

4 Methodische Entwicklungen

Die in dieser Arbeit entwickelte Methode zur Analyse der strukturellen Eigenschaften von Proteinkomplexen und der anschließenden Vorhersage von Bindungsstellen kann in drei einander folgende Schritte unterteilt werden:

1. Bestimmung physikalischer und chemischer Daten für die in der Brookhaven Proteindatenbank gespeicherten Proteinkomplexe auf der Basis der dreidimensionalen Strukturen der Proteine und anschließende Projektion der Daten auf die molekularen Oberflächen der Proteine.
2. Analyse der gesammelten Daten unter besonderer Berücksichtigung der Bindungseigenschaften von Protein-Protein-, Protein-DNA- und Protein-Ligand-Komplexen
3. Verwendung der Analyseergebnisse zur Vorhersage möglicher Bindungsstellen auf den molekularen Oberflächen von Proteinen

Der erste Schritt ist wiederum in mehrere Unterschritte aufgeteilt, die in Kapitel 4.1 einzeln beschrieben werden. Zuerst werden die Strukturdaten der Proteindatenbank gesichtet und auf Fehler überprüft. Dies schließt auch die Ergänzung fehlender Atompositionen in den Proteinstrukturen ein. Dazu wird das Programm CHARMM von Karplus und Mitarbeitern [95] verwendet. Nach Vervollständigung der Molekülstruktur und Aussondern von nicht geeigneten Proteinen und Proteinkomplexen werden im nächsten Schritt die physikalischen und chemischen Daten der Atome, wie z.B. Partialladungen, lipophile Fragmentwerte etc., bestimmt. Anschließend werden die molekularen Oberflächen berechnet und die in Kapitel 3.2 beschriebenen molekularen Eigenschaften auf diese Oberflächen projiziert. Darauf folgt die Bestimmung der Komplexbindungsbereiche. Zur weiteren Untersuchung werden schließlich die molekularen Oberflächen in definierte Teilbereiche aufgeteilt und für jede dieser Teiloberflächen die charakteristischen Merkmale zusammengefaßt.

In Kapitel 4.2 werden die Methoden, die bei der Analyse der gesammelten Daten verwendet werden, erläutert. Es werden Wasserstoff- und Disulfidbrücken zwischen den Proteinen bestimmt, die Wasserstoffakzeptoren- und Wasserstoffdonatoreneigenschaften an den Proteinoberflächen untersucht und die Verteilung der verschiedenen Aminosäuretypen an den Molekülоberflächen und die damit verbundenen molekularen Eigenschaften analysiert. Besonderes Augenmerk wird bei der Analyse auf die erkannten Bindungs-

bereiche von Protein-Protein-, Protein-DNA- und Protein-Ligand-Komplexen gelegt. Diese Bereiche werden auf ihre Größe, Form und Verteilung der molekularen Eigenschaften hin untersucht und die Ergebnisse den Untersuchungen der Gesamtoberfläche gegenübergestellt.

Die Analyseergebnisse bilden die Basis für die Entwicklung der neuen Methoden zur Vorhersage von möglichen Bindungsstellen. In Kapitel 4.3 und 6 werden Möglichkeiten zur Formulierung solcher Ansätze erläutert. Zuerst wird versucht, mittels einer einfachen Zielfunktion eine Bewertung der erzeugten Teiloberflächen hinsichtlich Bindungsfähigkeit in Proteinkomplexen zu erreichen. Es folgt die Beschreibung eines einfachen neuronalen Netzes, welches diese Aufgabe im Gegensatz zu der Zielfunktion zufriedenstellend lösen kann.

4.1 Automatisiertes Verfahren zur Analyse von Proteinstrukturen

4.1.1 Die Proteindatenbank

Grundlage der Untersuchungen der vorliegenden Arbeit ist die Kenntnis der vollständigen dreidimensionalen Struktur der betrachteten Moleküle. Die Koordinaten aller Atomschwerpunkte müssen bekannt sein. Für die Gewinnung dieser Daten gibt es mehrere Methoden. Die wichtigste experimentelle Methode zur Strukturaufklärung von Biopolymeren ist die Kristallstrukturanalyse. Über Röntgenbeugungsuntersuchungen an Proteinkristallen können die Positionen der Schweratome bestimmt werden. Bei kleinen Molekülen sind auch die Koordinaten der Wasserstoffatome mit dieser Methode zugänglich. Für Makromoleküle sind zusätzliche kostspielige Neutronenbeugungsexperimente zur Bestimmung der Wasserstoffpositionen nötig, worauf aber meistens verzichtet wird. In den letzten Jahren hat die Verwendung der 2D-Kernspinresonanzspektroskopie (NMR) zur Strukturaufklärung an Bedeutung gewonnen. Ein Vorteil der 2D-NMR-Spektroskopie ist, daß sie auch Informationen über die Positionen der Wasserstoffatome liefert. Die Ergebnisse experimenteller Strukturuntersuchungen werden in einigen zentralen Datenbanken gespeichert und so Forschern weltweit zugänglich gemacht. Die bedeutendste Strukturdatenbank für kleine Moleküle ist die *Cambridge Structural Database* (CSD) [112]. Sie enthält die Ergebnisse von Röntgenstrukturuntersuchungen an über 250000 kleinen organischen Molekülen. Die Strukturen von Biopolymeren wie Proteinen, Proteinkomplexen und Nukleinsäuren sind in der *Protein Data Bank* (PDB) abgelegt [5]. Zur Zeit verzeichnet die PDB über 17000 Einträge (Stand Januar 2001). Die

Zahl der jährlich neu bestimmten Proteinstrukturen steigt exponentiell an. Die Strukturen sind über das Internet oder auf CDROM vom National Institute of Standards and Technology (NIST) frei erhältlich [113]. Die in dieser Arbeit untersuchten Proteinstrukturen wurden dieser öffentlich zugänglichen Datenbank entnommen. Insgesamt standen so 10213 Strukturen (Stand Oktober 1999) für die Untersuchungen zur Verfügung.

4.1.2 Vorbereitung der Strukturdaten

Wie oben erwähnt sind die Koordinaten für Wasserstoffatome nicht in allen in der Proteindatenbank abgelegten Strukturen enthalten. Bei der Überprüfung der PDB-Einträge wurde deutlich, daß in vielen Strukturen auch einige Schweratome fehlen. Teilweise sind für bestimmte Atome auch mehrere mögliche Positionen angegeben. Für die weiteren Untersuchungen werden die Koordinaten aller Proteinatome benötigt. Es müssen folglich alle Molekülstrukturen vor der Weiterverarbeitung überprüft und gegebenenfalls ergänzt werden. Ein weiteres Problem stellt das nicht durchgängig einheitliche Format der Einträge dar. Bei einigen Strukturdaten weichen die Bezeichnungen von Atomtypen voneinander ab. Diese kleinen Abweichungen sind aber ein Problem für die automatische Bearbeitung und müssen beseitigt werden. Die Analyse und Verarbeitung der Ausgangsdaten wird in mehrere Schritte unterteilt.

Im ersten Schritt werden das Format der Datenbankeinträge überprüft und etwaige Abweichungen vom Standardformat beseitigt. Hierbei wird besonders auf die Erkennung der richtigen Atomtypen und Aminosäuren geachtet. Danach wird überprüft, ob für alle aufgeführten Atome Koordinaten vorhanden sind. Sind für ein Atom mehrere Koordinaten angegeben, so ist zusätzlich für jede Position eine Angabe über die Besetzungswahrscheinlichkeiten der verschiedenen Positionen vorhanden. Für die weiteren Berechnungen wird die Atomposition mit der höchsten Besetzungswahrscheinlichkeit verwendet. Wenn nicht alle Atomtypen fehlerlos erkannt werden, wird der gesamte Datenbankeintrag verworfen. Auch Einträge, die nur DNA-Moleküle oder Kohlenhydrate, jedoch keine Proteinstrukturen enthalten, werden aussortiert. Nach Überprüfung des Datenformates wird die Proteinkomplexstruktur analysiert und in ihre Untereinheiten aufgeteilt. Alle nicht über kovalente Atombindungen verbundenen Moleküle werden zur weiteren Bearbeitung in einzelnen Dateien getrennt abgespeichert.

Zur Ergänzung fehlender Atompositionen wird das Programm CHARMM Version 24 [95] verwendet. CHARMM stellt Algorithmen zur Simulation von Makromolekülen zur Verfügung und enthält Kraftfelder zur Behandlung von Proteinen, Nukleinsäuren und

Kohlehydraten. Moleküle, deren Atome durch diese Kraftfelder nicht abgedeckt sind, wie z.B. nichtpeptidische Liganden, kann CHARMM nicht verarbeiten. Sie werden daher ignoriert. Problematisch sind Proteinstrukturen mit Aminosäurederivaten oder nicht-proteinogenen Aminosäuren in der Polypeptidkette. Datenbankeinträge mit solchen Proteinstrukturen kann das hier vorgestellte Verfahren nicht weiter bearbeiten, und sie werden verworfen.

Nach Konvertierung der PDB-Daten in das Dateiformat von CHARMM werden die fehlenden Wasserstoff- und Schweratome hinzugefügt. Dazu benötigt CHARMM explizite Aussagen über den Protonierungsgrad der beteiligten Atomgruppen. Wenn der Protonierungszustand in der Proteinstruktur nicht angegeben ist, werden folgende Annahmen gemacht:

1. Das Phosphatrückgrat von Nukleinsäuren ist vollständig deprotoniert.
2. Arginin und Lysin sind einfach positiv, Glutaminsäure und Asparaginsäure einfach negativ geladen.
3. Histidin wird neutral behandelt. Wenn Histidin an der Bindung zu Metallionen beteiligt ist, wird das nicht an dieser Bindung beteiligte Stickstoffatom protoniert.
4. Anhand Abstandsberechnungen wird überprüft, ob Cystein Disulfidbrücken bildet. Die Schwefelatome der Cysteinseitenkette sind auch oft an Bindungen zu Metallionen beteiligt. In beiden Fällen werden keine Wasserstoffatome an die Schwefelatome hinzugefügt.

Nach der Plazierung der fehlenden Atome wird eine Energieminimierung der neuen Atompositionen mit dem *adopted-basis-set-Newton-Raphson*-Verfahren in CHARMM vorgenommen. Alle anderen Atome, deren Koordinaten im Datenbankeintrag schon vorhanden waren, werden dabei durch Zwangskräfte (*Fix Constraints*) auf ihren vorgegebenen Positionen fixiert. Die mit CHARMM bearbeiteten Strukturen werden abschließend noch einmal überprüft und in das PDB-Datenformat zurück konvertiert. Ingesamt können mit diesem Verfahren für 82% der Proteine in der *Protein Data Bank* die vollständigen Strukturinformationen erzeugt werden (siehe Kapitel 5).

4.1.3 Zuordnung der physikochemischen Daten und Parameter

Vor der Energieminimierung der hinzugefügten Atompositionen weist CHARMM jedem Atom Partialladungen zu und markiert Wasserstoffdonatoren und -akzeptoren. Diese

Informationen werden für die Berechnung der physikalischen und chemischen Eigenschaften und die anschließende Projektion auf die molekularen Oberflächen benötigt und abgespeichert. Die Berechnung der lokalen Lipophilie basiert, wie bereits beschrieben, auf den atomaren Lipophilie-Partialwerten von Crippen. Die Zuordnung der Proteinatome zu den 120 Crippen-Strukturtypen geschieht mit dem von Jäger entwickelten Programm MOLFESD [114]. Die Fluktuationen der Atompositionen sind in den originalen PDB-Dateien in Form der B-Faktoren enthalten. Bei theoretischen Modellen oder durch NMR gelösten Strukturen sind die B-Faktoren nicht vorhanden, und somit kann die Flexibilität des Proteins nicht untersucht werden. Eine zusätzliche Untersuchung der Flexibilität mit einer MD-Simulation (Kapitel 3.2.6) wäre für die große Anzahl von Proteinen zu aufwendig.

4.1.4 Berechnung der molekularen Oberfläche und Projektion der Eigenschaften

Die Berechnung der molekularen Oberfläche der Proteine erfolgt mit dem neu entwickelten Programm FUMEE, welches die in Kapitel 3.1.4 vorgestellte Methode implementiert. Gleichzeitig werden die Aminosäuren des Proteins an der Oberfläche gezählt und der jeweilige Anteil der Aminosäure an der Proteinoberfläche bestimmt. Über diesen Oberflächenanteil wird definiert, welche Aminosäure zur Proteinoberfläche und welche zum Proteininneren gehört. Ebenso werden die Atome an der Oberfläche gesucht und für weitere Untersuchungen markiert.

Die physikalischen und chemischen Eigenschaften der Proteine werden mit den in Kapitel 3.2 erläuterten Algorithmen berechnet und auf die molekularen Oberflächen projiziert. Tabelle 4.1 listet die berechneten Eigenschaften und die dazu verwendeten Programme auf. Das Programm QUALEN 2 ist eine Weiterentwicklung der Programme QUAL [7] und QUALEN [65]. In QUALEN 2 wurde der Programmablauf optimiert, so daß die Berechnung der molekularen Eigenschaften schneller als bei seinen beiden Vorgängern abläuft. Ohne diese Optimierungen wäre die Berechnung der Eigenschaften von tausenden Proteinen zu zeitaufwendig, und die Untersuchung müßte auf eine geringere Anzahl von Komplexen reduziert werden. Die lokale Wasserstoffakzeptoren- bzw. Wasserstoff-donatorendichte und die Tiefe der Taschen und Spalten in der Oberfläche wird mit HBDENS bzw. DEEPQUAL berechnet, beide Programme wurden entsprechend der in Kapitel 3.2.3.2 und 3.2.5 erläuterten Methoden implementiert. Die Oberflächen-krümmungen werden mit dem Programm GLOBCURV von Exner [64] berechnet. Zur Bestimmung der Bindungsbereiche von Proteinkomplexen wird zusätzlich das Programm

INTERFACING verwendet. Es unterteilt die Oberfläche in bindende und nichtbindende Bereiche entsprechend des im Kapitel 3.2.7 beschriebenen Abstandkriteriums zu den Komplexpartnern. Es unterscheidet dabei zwischen Protein-, DNA- und Ligandbindungsbereichen der Proteine.

Tabelle 4.1: Programme zur Berechnung der molekularen Eigenschaften.

Eigenschaft		Verwendetes Programm
Elektrostatisches Potential	(Kapitel 3.2.1)	QUALEN 2
Lokale Lipophilie	(Kapitel 3.2.2)	QUALEN 2
Molekulare Flexibilität	(Kapitel 3.2.6)	QUALEN 2
Wasserstoffakzeptoren- bzw. Donatorendichte	(Kapitel 3.2.3.2)	HBDENS
Tiefeninformation	(Kapitel 3.2.5)	DEEPQUAL
Oberflächenkrümmung	(Kapitel 3.2.4)	GLOBCURV
Intermolekulare Abstände - Bindungsregionen	(Kapitel 3.2.7)	INTERFACING

Bei den Berechnungen werden folgende Parameter verwendet:

- Berechnung der Molekularen Oberfläche (Kapitel 3.1.4):
  Radius der Probekugel: 1,4 Å
  Punktdichte der Oberfläche: 7,0 Punkte pro$Å^2$
- Elektrostatisches Potential (Gleichung 3.5):
  Cutoff: 10,0 Å
- Lokale Lipophilie (Gleichung 3.9):
  Parameter a: 1,0
  Parameter b: 2,5
- Molekulare Flexibilität (Gleichung 3.9):
  Parameter a: 1,0
  Parameter b: 2,5
- Wasserstoffakzeptoren- und Wasserstoffdonatorendichte (Gleichung 3.10):
  Cutoff-Radius: 4,0 Å
- Tiefe von Taschen und Spalten in der Proteinoberfläche (Kapitel 3.2.5):
  Radius der abdeckenden Probekugel B: 6,0 Å
- Oberflächenkrümmung (Kapitel 3.2.4):
  Cutoff-Radius: 6,0 Å
- Intermolekularer Abstand – Bestimmung von Bindungsregionen (Kapitel 3.2.7):
  Maximaler Abstand innerhalb einer Bindungsregion: 1,5 Å

4.1.5 Aufteilung der molekularen Oberfläche in Teilbereiche

Zur weiteren Analyse der Oberflächeneigenschaften wird die molekulare Oberfläche in kleinere Teilbereiche unterteilt. Heiden [115] entwickelte ein auf *Fuzzy Logic* basierendes Verfahren zur Segmentierung von Oberflächen anhand der auf ihnen projizierten Eigenschaften. Dabei wird versucht, Teiloberflächen aus benachbarten Oberflächenpunkten mit möglichst ähnlichen Eigenschaften zu bilden. Dieser Algorithmus generiert jedoch Oberflächensegmente, die sich in ihrer Form und Größe stark unterscheiden können, was bei der späteren Analyse der lokalen Oberflächeneigenschaften und angestrebten Vorhersage von Bindungsbereichen zu Problemen führen kann. Von Exner [64,116,117] wurde eine ebenfalls auf *Fuzzy Logic* basierende Methode zur Segmentierung molekularer Oberflächen entwickelt, die jedoch Teiloberflächen sehr ähnlicher Form und Größe erzeugt. Das Verfahren ermittelt zuerst für jeden Oberflächenpunkt anhand von gegebenen Funktionen (*Membership Functions*) die jeweilige Zugehörigkeit zu mehreren Eigenschaftsklassen. Bei der Einteilung nach z.B. elektrostatischem Potential gibt es fünf Klassen: stark negativ, negativ, neutral, positiv, stark positiv. Anschließend werden lokale Maxima der Zugehörigkeiten zu den einzelnen Klassen bestimmt. Die Teiloberflächen werden durch ringförmiges Wachstum ausgehend von diesen Extrempunkten gebildet, wobei eine zu große Unähnlichkeit der Eigenschaften zum Extrempunkt oder das Erreichen einer Maximalgröße dieses Wachstum beendet. Teiloberflächen, die kleiner als eine gegebene Minimalgröße sind, werden verworfen. Die so erzeugten Segmente der molekularen Oberfläche überlappen teilweise und haben annähernd kreisförmige Gestalt.

In dieser Arbeit wird als Aufteilungskriterium die Wasserstoffakzeptoren- und Wasserstoffdonatorendichte benutzt (siehe Kapitel 3.2.3.2). Die verwendete Zugehörigkeitsfunktion μ (Diagramm 4.1) teilt die Oberfläche in Teilbereiche mit hoher bzw. niedriger Dichte ein [116]. Die Minimal- und Maximalgröße der Oberflächensegmente beträgt 50 $Å^2$ bzw. 250 $Å^2$. So entsprechen die Größen der Segmente ungefähr der Fläche, die zwei bis drei Aminosäuren an der molekularen Oberfläche einnehmen.

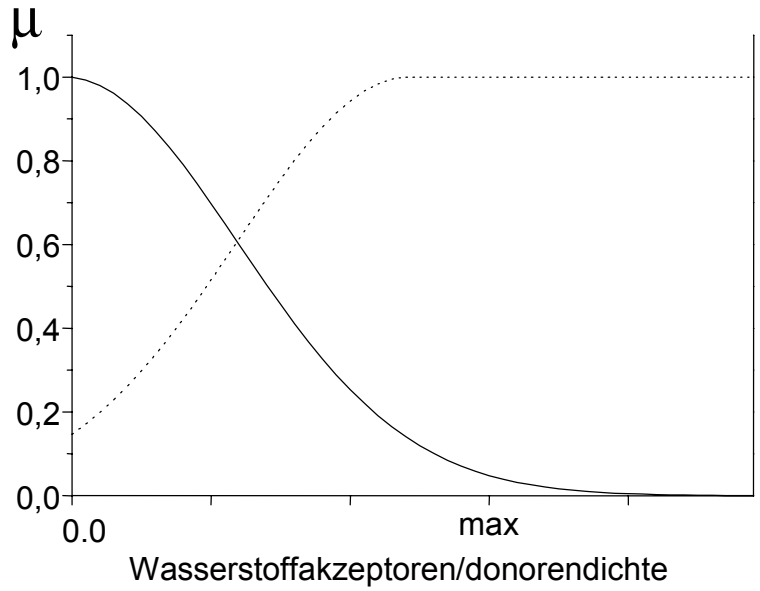


Diagramm 4.1: Zugehörigkeitsfunktion zur Einteilung der molekularen Oberfläche anhand der Wasserstoffakzeptoren- und -donatorendichte: (—) niedrige Dichte, (···) hohe Dichte. Der Maximalwert max beträgt 0,15 $Å^{-2}$.

Für jedes Oberflächensegment werden die im vorhergehendem Schritt berechneten Eigenschaften über alle Oberflächenpunkte des Segmentes gemittelt:

$$X_{OS} = \sum_i X_i \cdot \frac{A_i}{A_{OS}} \tag{4.1}$$

mit

X_{OS}: gemittelte Eigenschaft des gesamten Oberflächensegmentes

X_i: Eigenschaft des Oberflächenpunktes i

A_{OS}: Fläche des gesamten Oberflächensegmentes [$Å^2$]

A_i: Fläche des Oberflächenpunktes i [$Å^2$]

Zusätzlich wird für jede Teiloberfläche der Oberflächenbereich bestimmt, der an Bindungen zu anderen Molekülen beteiligt ist (siehe Kapitel 3.2.7). Dabei werden drei Arten von Bindungsbereichen unterschieden: Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereiche. Die Zugehörigkeit zu der jeweiligen Bindungsregion (Bindungsanteil der Teiloberfläche) wird über den Quotienten aus Bindungsfläche und Gesamtfläche der Teiloberfläche definiert. Darauf basierend werden alle Teiloberflächen einer von vier Klassen zugeordnet. Dies sind die drei Arten von Bindungsbereichen

(Protein, DNA und Ligand) und die Gruppe der nichtbindenden Oberflächenbereiche. Die Zuordnung geschieht nach folgenden Regeln: Wenn die Teiloberfläche in einem Bindungsbereich der molekularen Oberfläche liegt, wird sie der Bindungsbereichart mit dem größten Anteil an der Teiloberfläche zugeordnet, sofern diese Bindungsfläche mindestens 20% der Gesamtfläche des Oberflächensegmentes ausmacht. Dieser kleine Grenzwert ist nötig, damit auch Teiloberflächen, die nur am Rand von Bindungsbereichen oder in sehr kleinen Bindungsbereichen (z.B. Ligandbindungsstellen) liegen, als Bindungsflächen eingeordnet und behandelt werden. Wird der Grenzwert von 20% Bindungsfläche nicht erreicht, oder die Teiloberfläche liegt in keinem Bindungsbereich der molekularen Oberfläche, wird sie als nichtbindende Teiloberfläche klassifiziert. Die über die Teiloberfläche gemittelten Eigenschaften und die Zugehörigkeit zu den Bindungsbereicharten bilden die Basis der anschließenden Analysen.

4.2 Analyse der molekularen Eigenschaften

4.2.1 Einteilung der Proteinstrukturen in verschiedene Datensätze

Wie oben erwähnt umfaßt die *Protein Data Bank* die Strukturdaten von mehreren tausend Proteinen und Proteinkomplexen. Aufgrund dieser großen Anzahl von Strukturen und den teilweise sehr aufwendigen Berechnungen, die zur Analyse der Proteineigenschaften nötig sind, wurden in der Vergangenheit viele Untersuchungen von Proteineigenschaften nur an Teilmengen der Datenbank durchgeführt. Dazu wurden reduzierte Datensätze erstellt, die mit möglichst wenigen Proteinstrukturen die strukturellen Eigenschaften aller Proteine der gesamten Datenbank abdecken. Um eine Reduktion der Gesamtdaten mit möglichst wenig Informationsverlust zu erreichen, werden Proteinstrukturen mit möglichst hoher Diversität ausgewählt, d.h. redundante und sehr ähnliche Strukturen ausgesondert. Solch ein Datensatz wurde z.B. von Tsai et al. [26,118] erstellt und bei der Analyse von Protein-Protein-Komplexen verwendet. Ein anderer Ansatz ist die *PDB-Select*-Liste von Hobohm et al. [119,120]. Diese Liste enthält ca. 500 Proteine mit einer maximalen Sequenzidentität von 25% und einer Auflösung der Kristallstruktur von unter 3,0 Å. Viele Autoren beschränken ihre Untersuchungen auch auf bestimmte Typen von Proteinkomplexen und reduzieren dadurch die Anzahl der zu untersuchenden Strukturen. Es gibt z.B. Untersuchungen an ausgewählten Protein-DNA- [37,121], Enzym-Inhibitor- [19,20,22,24, 25,28], oder Antigen-Antikörper-Komplexen [59,122,123].

Für die Auswertungen in dieser Arbeit werden vier Datensätze definiert. Der erste Datensatz enthält alle verwendeten Strukturen. Er wird als Datensatz A bezeichnet. Ein reduzierter Datensatz B wird mit der *PDB-Select*-Liste (Stand März 2001) von Hobohm et al. erstellt [119]. Für die Untersuchung der speziellen Bindungseigenschaften von Enzym-Inhibitor- und Antikörper-Antigen-Komplexen werden noch entsprechende Datensätze C und D erstellt. Dazu werden die Datenbankeinträge nach adäquaten Stichwörtern durchsucht. Mit den Stichwörtern „*antigen*“, „*antibody*“, „*immune*“ und „*immuno-globuline*“ und einer anschließenden manuellen Verifizierung der Suchergebnisse ergibt sich der Datensatz C mit Antigen-Antikörper-Komplexen. Für den Enzym-Inhibitor-Datensatz D wird eine Suche nach dem Stichwort „*inhibitor*“ gestartet. Dabei wird darauf geachtet, daß nur Proteinkomplexe mit proteinogenen Inhibitoren in den Datensatz aufgenommen werden. Um Proteinkomplexe mit niedermolekularen Inhibitoren auszuschließen, werden zusätzlich nur Peptide mit mindestens neun Aminosäuren als Inhibitoren akzeptiert. Die Zusammensetzung der so definierten Datensätze ist in Tabelle 5.5 im Kapitel 5.1.3 aufgelistet.

4.2.2 Intermolekulare Wasserstoffbrückenbindungen

In den Bindungsbereichen der Proteinkomplexe bilden sich oft Wasserstoffbrücken und/oder Salzbrücken zwischen den Komplexpartnern. Diese intermolekularen Wechselwirkungen haben Einfluß sowohl auf die Stärke als auch die Spezifität der Komplexbindung. Deshalb ist es wichtig, die Komplexbindungsbereiche nach Wasserstoffbrückenbindungen und Salzbrücken abzusuchen. Das Programm HBOND überprüft alle möglichen Atom-Atom-Kontakte innerhalb der Bindungsregion der Molekülkomplexe auf mögliche Brückenbindungen. Durch die Bearbeitung der Proteinstrukturen mit CHARMM ist für jedes Atom bekannt, ob es als Wasserstoffakzeptor oder Donator in Frage kommt. Sobald ein Donator/Akzeptor-Paar X–H···Y die folgenden Kriterien erfüllt, wird es als Wasserstoffbrücke markiert: Der Abstand zwischen H und Y beträgt maximal 3,0 Å, die maximale Distanz von X nach Y ist 3,2 Å und der Winkel X–H···Y beträgt maximal 90° [124]. Dabei gibt der genannte Winkel die Abweichung von der gestreckten Anordnung der drei Atome an, d.h. ein Winkel von 0° entspricht einer linearen Wasserstoffbrücke. Zusätzlich zur Anzahl der gebildeten Wasserstoffbrücken pro Bindungsbereich wird auch die Anzahl der Wasserstoffakzeptoren und Wasserstoffdonatoren an der Gesamtoberfläche der Proteine und innerhalb der Bindungsbereiche gezählt, so daß die Anzahl der inter-

molekularen Wasserstoffbrückenbindungen mit der Menge der Wasserstoffakzeptoren bzw. Donatoren im Bindungsbereich verglichen werden kann.

4.2.3 Wasserstoffakzeptoren und Wasserstoffdonatoren

Aus der Anzahl der Wasserstoffakzeptoren und Wasserstoffdonatoren, die sich an der Oberfläche der Proteine befinden, und dem Flächeninhalt der gesamten Proteinoberfläche bzw. den Bindungsbereichen werden die entsprechenden Akzeptoren- und Donatorendichten berechnet. Im Gegensatz zu den lokalen Akzeptoren- bzw. Donatorendichten für jeden Oberflächenpunkt (siehe Kapitel 3.2.3.2) wird hier ein Dichtewert über die gesamte Fläche berechnet. Bei den Bindungsbereichen werden Akzeptoren bzw. Donatoren, die sich am Rand des Bindungsbereiches befinden, besonders behandelt. Äquivalent der Berechnung der lokalen Dichten wird der Anteil des betreffenden Atoms im Bindungsbereich bestimmt und nur dieser berücksichtigt.

4.2.4 Radiale Verteilungsfunktion der Wasserstoffdonatoren und Akzeptoren

Die Untersuchung der Anordnung der Wasserstoffakzeptoren und -donatoren an der Proteinaußenseite erfolgt mit Hilfe von radialen Paarverteilungsfunktionen. Die radialen Verteilungsfunktionen der Abstände zwischen Akzeptoren und Donatoren werden, wie in Kapitel 3.2.3.3 beschrieben, berechnet. Insgesamt werden vier verschiedene radiale Paarverteilungsfunktionen berechnet und analysiert: Akzeptor-Akzeptor, Akzeptor-Donator, Donator-Akzeptor und Donator-Donator. Dabei bedeutet z.B. Akzeptor-Donator, daß um jedes Wasserstoffakzeptoratom an der Proteinoberfläche die radiale Verteilung von Wasserstoffdonatoratomen bestimmt wird. Die radialen Verteilungsfunktionen um jeden Akzeptor bzw. Donator eines Proteins werden jeweils zu einer mittleren Verteilungsfunktion zusammengefaßt und dann nochmals über alle Proteine gemittelt, um charakteristische Muster der Akzeptor-Donator-Verteilung herauszuarbeiten. Bei der Berechnung der Verteilungsfunktionen werden zwei Fälle unterschieden: Zum einen werden alle möglichen Atompaarungen berücksichtigt. Im anderen Fall müssen die betrachteten Wasserstoffakzeptoren bzw. -donatoren zu unterschiedlichen Aminosäuren gehören. Für beide Fälle werden die radialen Verteilungsfunktionen an der molekularen Gesamtoberfläche und in den Bindungsbereichen berechnet und gegenübergestellt.

4.2.5 Aminosäurenzusammensetzung der Proteine und Proteinoberflächen

Die Analyse der Proteinstrukturen umfaßt eine Statistik der Größe (Atom- und Aminosäurenanzahl pro Protein, bzw. Molekulargewicht), des Flächeninhaltes der erzeugten molekularen Oberfläche und des Volumens der Proteine auf der Basis dieser Oberfläche. Ebenso wird die Größe der verschiedenen Bindungsbereiche in den Protein-Komplexen untersucht. Die Aminosäurenzusammensetzung der Proteine, Oberflächen und Bindungsbereiche wird über die relativen Häufigkeiten der 20 proteinogenen Aminosäuren berechnet. Die relative Häufigkeit n_{rel} der Aminosäuren im gesamten Protein ergibt sich aus der Anzahl der Aminosäuren eines Typs geteilt durch die Gesamtanzahl aller Aminosäuren im Protein. Die relative Häufigkeit a_{rel} der einzelnen Aminosäurentypen an der Proteinoberfläche wird in dieser Arbeit über ihren Anteil an der Gesamtoberfläche bestimmt.

$$a_{rel,X} = \frac{\sum_i A_{X,i}}{A_{gesamt}} \tag{4.2}$$

mit

$a_{rel,X}$: Relative Häufigkeit der Aminosäure X an der Proteinoberfläche

A_{gesamt}: Oberflächeninhalt des Proteins [$Å^2$]

$A_{X,i}$: Teil der Proteinoberfläche, welcher der Aminosäure i des Typs X zugeordnet ist [$Å^2$]

Neben dem Anteil, den die einzelnen Aminosäuretypen an der Proteinoberfläche besitzen, wird auch der Mittelwert der Fläche $A_{X,i}$, die einer Aminosäure eines bestimmten Typs X an der molekularen Oberfläche zugeordnet wird, bestimmt. Zu Vergleichszwecken werden zusätzlich die Strukturen von 20 Tripeptiden (Glycin-X-Glycin, X = 20 proteinogene Aminosäuren) mit Hilfe des *Molecular-Modeling*-Programmpakets SYBYL von Tripos [125] erzeugt und mit dem Kraftfeld von SYBYL minimiert. Anschließend wird die molekulare Oberfläche der Tripeptide berechnet und der Anteil der Aminosäure X an der Oberfläche bestimmt. Mit dieser maximalen Fläche, die eine Aminosäure in einer nicht gefalteten Polypeptidkette besitzen kann, wird der Oberflächenverlust $f_{rel,i}$ der einzelnen Aminosäuren im Protein berechnet. Der Flächenverlust kommt durch die Faltung der Peptidkette zum globulären Protein zustande. Je niedriger der Flächenverlust, desto weiter ragt die Aminosäure aus dem Protein heraus. In einigen Untersuchungen wird der relative

Oberflächenverlust $f_{rel,i}$ (*Fractional Surface Area Loss* oder *Fractional Accessibility*) auch als Maß für die Hydrophobie von Aminosäuren herangezogen [35,126].

$$f_{rel,i} = \frac{A_i^0 - A_i}{A_i^0} \tag{4.3}$$

mit

$f_{rel,i}$: Relativer Oberflächenverlust der Aminosäure i

A_i^0: Oberfläche der einzelnen Aminosäure i im Tripeptid Glycin-X-Glycin [$Å^2$]

A_i: Fläche der einzelnen Aminosäure i an der Proteinoberfläche [$Å^2$]

Die proteinogenen Aminosäuren unterscheiden sich untereinander nur in der Struktur ihrer Seitenketten. Die restlichen Atome, welche das Proteinrückgrat bilden, sind in allen Aminosäuren mit Ausnahme von Prolin identisch. Dieser gemeinsame Molekülteil besitzt jeweils einen Wasserstoffdonator, einen Wasserstoffakzeptor und hat einen hydrophilen Charakter. Je nach Anordnung der Aminosäuren an der Proteinoberfläche können also auch Aminosäuren mit hydrophoben Seitenketten, wie z.B. Leucin, hydrophile Wechselwirkungen mit anderen Molekülen ausbilden. Um diesen Effekt zu untersuchen, wird sowohl die Oberfläche der gesamten Aminosäure als auch nur die Oberfläche der Seitenkette bestimmt und ausgewertet.

4.2.6 Aminosäurenkontakte in Protein-Protein-Bindungsbereichen

In den Bindungsregionen der Protein-Protein-Komplexe treten die Aminosäuren beider Komplexpartner auf vielfältige Weise in Kontakt. Neben Wasserstoff-, Salz- und Disulfidbrückenbindungen gibt es auch viele hydrophobe Wechselwirkungen zwischen den Aminosäuren beider Proteine. Um diese Wechselwirkungen zu untersuchen, werden die Größen der Bindungsflächen zwischen den 20 proteinogenen Aminosäuren im Bindungsbereich bestimmt und in einer (20,20)-Matrix (siehe Gleichung 4.4) aufsummiert [29]. Für jede mögliche Aminosäurenkombination dieser Aminosäuren wird ein Matrixelement belegt. Die gemeinsame Bindungsfläche $A_{Kontakt}^{X_i-Y_j}$ eines Aminosäurenkontaktpaares ergibt sich aus dem Mittelwert der Kontaktoberflächen $A_{Kontakt}^{X_i|Y_j}$ beider Aminosäuren zueinander (siehe Abbildung 4.1). Für die Untersuchung werden die Matrizen aller Protein-Protein-Bindungsregionen zu einer mittleren Kontaktmatrix zusammengefaßt.

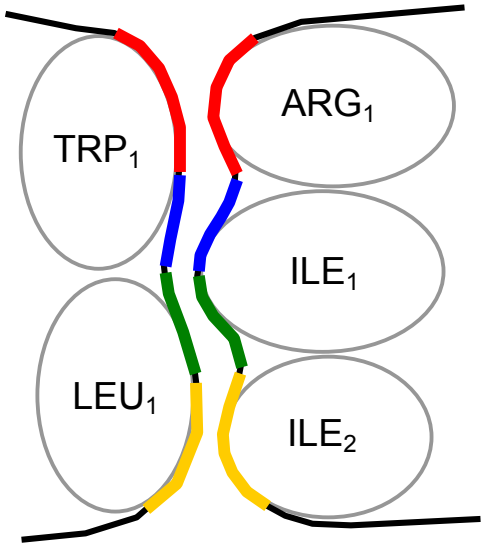


Abb. 4.1: Molekulare Oberfläche im Bindungsbereich eines Protein-Protein-Komplexes. Die Kontaktoberflächen zwischen den Aminosäuren der beiden Proteine sind farblich hervorgehoben:

rot: Kontaktoberflächen $A_{Kontakt}^{TRP_1|ARG_1}$ und $A_{Kontakt}^{ARG_1|TRP_1}$ zwischen TRP_1 und ARG_1

blau: Kontaktoberflächen $A_{Kontakt}^{TRP_1|ILE_1}$ und $A_{Kontakt}^{ILE_1|TRP_1}$ zwischen TRP_1 und ILE_1

grün: Kontaktoberflächen $A_{Kontakt}^{LEU_1|ILE_1}$ und $A_{Kontakt}^{ILE_1|LEU_1}$ zwischen LEU_1 und ILE_1

gelb: Kontaktoberflächen $A_{Kontakt}^{LEU_1|ILE_2}$ und $A_{Kontakt}^{ILE_2|LEU_1}$ zwischen LEU_1 und ILE_2.

Im vorhergehenden Kapitel 4.2.5 wird beschrieben, wie die relative Häufigkeit der Aminosäuren an der Proteinoberfläche berechnet wird. Unter der Annahme, daß alle Aminosäuren mit der gleichen Wahrscheinlichkeit untereinander in Wechselwirkung treten, kann eine theoretische Kontaktmatrix berechnet und mit der tatsächlich beobachteten verglichen werden. Die einzelnen Elemente der theoretischen Kontaktmatrix ergeben sich aus dem Produkt der jeweiligen relativen Häufigkeiten $a_{rel,X}$ der beiden betrachteten Aminosäurentypen. Die Differenz der beobachteten Kontaktmatrix und der numerisch berechneten Matrix sollte dann bevorzugte oder benachteiligte Wechselwirkungen zwischen Aminosäurenpaaren zeigen. Die Aminosäurenkontaktmatrix der Protein-Protein-Bindungsregionen wird wie folgt berechnet:

$$\mathbf{M}_{\text{Kontakt,beob.}} = \begin{pmatrix} \dfrac{\sum_i \sum_j A_{\text{Kontakt}}^{\text{LYS}_i-\text{LYS}_j}}{A_{\text{Kontakt,ges.}}} & \cdots & \dfrac{\sum_i \sum_j A_{\text{Kontakt}}^{\text{TRP}_i-\text{LYS}_j}}{A_{\text{Kontakt,ges.}}} \\ \vdots & \ddots & \vdots \\ \dfrac{\sum_i \sum_j A_{\text{Kontakt}}^{\text{LYS}_i-\text{TRP}_j}}{A_{\text{Kontakt,ges.}}} & \cdots & \dfrac{\sum_i \sum_j A_{\text{Kontakt}}^{\text{TRP}_i-\text{TRP}_j}}{A_{\text{Kontakt,ges.}}} \end{pmatrix} \tag{4.4}$$

mit

$$A_{\text{Kontakt}}^{X_i-Y_j} = \tfrac{1}{2}\left(A_{\text{Kontakt}}^{X_i|Y_j} + A_{\text{Kontakt}}^{Y_j|X_i}\right)$$

und

$\mathbf{M}_{\text{Kontakt,beob}}$: Matrix der Oberflächenkontakte zwischen den 20 proteinogenen Aminosäuren in der Bindungsregion von Protein-Protein-Komplexen

$A_{\text{Kontakt}}^{X_i|Y_j}$: Kontaktfläche [Å^2] der Aminosäure i des Typs X mit der Aminosäure j des Typs Y. Die Aminosäuren i und j befinden sich in unterschiedlichen Proteinen

$A_{\text{Kontakt}}^{X_i-Y_j}$: Mittlere Kontaktfläche zwischen den Aminosäuren i und j [Å^2]

$A_{\text{Kontakt,ges.}}$: Gesamtkontaktfläche zwischen den Proteinen [Å^2]

Die theoretische Kontaktmatrix lautet:

$$\mathbf{M}_{\text{Kontakt,theoretisch}} = \begin{pmatrix} a_{\text{rel,LYS}} \cdot a_{\text{rel,LYS}} & \cdots & a_{\text{rel,TRP}} \cdot a_{\text{rel,LYS}} \\ \vdots & \ddots & \vdots \\ a_{\text{rel,TRP}} \cdot a_{\text{rel,LYS}} & \cdots & a_{\text{rel,TRP}} \cdot a_{\text{rel,TRP}} \end{pmatrix} \tag{4.5}$$

mit

$\mathbf{M}_{\text{Kontakt,theoretisch}}$: Theoretische Matrix der Oberflächenkontakte

$a_{\text{rel,X}}$: Relative Häufigkeit der Aminosäure X auf der Proteinoberfläche

4.2.7 Analyse der Proteinoberflächen

Für die Untersuchung der auf die molekulare Oberfläche projizierten Eigenschaften (elektrostatisches Potential, lokale Lipophilie, Oberflächenkrümmung, etc.) werden die Mittelwerte der Eigenschaften über alle Teiloberflächen ermittelt (siehe Kapitel 4.1.4 und 4.1.5) und dann die Eigenschaftswerte an der Gesamtoberfläche mit den Werten in den Bindungsbereichoberflächen der Proteine (Protein-Protein, Protein-DNA und Protein-Ligand) verglichen. Neben der Analyse der Mittelwerte werden auch die Verteilung und Streuung der Eigenschaftswerte berechnet und untersucht. Dabei wird wiederum zwischen der Gesamtoberfläche und den Bindungsbereichen der Proteine unterschieden. Die Ergebnisse dieser Untersuchungen sind die Basis für die Entwicklung von Methoden zur Vorhersage von Bindungsbereichen der Proteine.

4.3 Methoden zur Vorhersage von Bindungsbereichen in Proteinen

Neben der Analyse der auf die Proteinoberflächen projizierten molekularen Eigenschaften und den Bindungseigenschaften der Proteinkomplexe ist die Entwicklung eines Algorithmus zur Vorhersage möglicher Bindungsbereiche der Proteinoberflächen Ziel dieser Arbeit. Der Algorithmus soll in der Lage sein, Oberflächenbereiche, die an Komplexbindungen zu anderen Molekülen beteiligt sein können, anhand ihrer Eigenschaften zu identifizieren. Dabei soll zwischen Protein-, DNA- und Ligandbindungsbereichen der Proteine unterschieden werden. Der Algorithmus greift zur Lösung dieser Aufgabe auf die in Kapitel 4.1 vorgestellten Methoden zurück und berechnet für die zu untersuchenden Proteine die molekulare Oberfläche und Eigenschaften. Anschließend werden die Oberflächen, wie dort beschrieben, in Teiloberflächen unterteilt und die Mittelwerte der auf die Teiloberflächen projizierten Eigenschaften berechnet. Diese Werte sind die Eingabedaten für den Vorhersagealgorithmus. In der vorliegenden Arbeit werden zwei verschiedene Methoden zur Erkennung der Bindungsbereiche angewendet: Die erste Methode basiert auf der Annahme, daß die Zugehörigkeit einer Teiloberfläche zu den verschiedenen Arten von Bindungsbereichen durch eine Linearkombination der gemittelten Eigenschaftswerte der Teiloberfläche berechnet werden kann. Für die Parametrisierung dieser Zielfunktion stehen die berechneten und ausgewerteten molekularen Eigenschaften aller Teiloberflächen des Gesamtdatensatzes zur Verfügung. Für jeden der drei verschiedenen Bindungsbereiche (Protein-, DNA- und Ligandbindungsstelle) und die nichtbindende Oberfläche wird eine Zielfunktion aufgestellt und die Koeffizienten durch

multilineare Regression bestimmt. Die zweite Methode verwendet ein neuronales Netz zur Vorhersage. Neuronale Netze können auf die Erkennung von spezifischen Mustern, d.h. auch auf die Erkennung von möglichen Bindungsbereichen, trainiert werden. Als Eingabemuster dienen wiederum die Mittelwerte der auf die Teiloberflächen projizierten physikalischen und chemischen Eigenschaften, und die Ausgabeneuronen des Netzwerkes repräsentieren den Anteil der Teiloberflächen an möglichen Bindungsstellen. Zum Training und Test des Netzwerkes werden die Teiloberflächen aller untersuchten Proteine verwendet.

Der genaue Aufbau bzw. die Parameter der Zielfunktionen und des neuronalen Netzwerkes hängen von den Ergebnissen der vorhergehenden Untersuchungen ab (Analyse der auf die Proteinoberflächen projizierten molekularen Eigenschaften), deshalb werden beide Methoden erst in Kapitel 6 ausführlich vorgestellt und erläutert. Dort werden auch die Vorhersageleistungen der beiden Ansätze präsentiert und diskutiert.

5 Ergebnisse

Die Untersuchungen basieren auf den Einträgen der Proteindatenbank vom Oktober 1999. Von den 10290 verzeichneten Einträgen enthalten 724 (7%) keine Proteine und wurden nicht bearbeitet. 1745 weitere Strukturen wurden während der Aufbereitung der Ausgangsdaten aussortiert. Der Hauptteil der ausgesonderten Proteinstrukturen konnte mit CHARMM nicht zufriedenstellend verarbeitet werden (siehe Kapitel 4.1.2). Die restlichen Strukturen durchliefen die vorgestellte Methode (Kapitel 4) problemlos, so daß 7821 Protein- und Proteinkomplexstrukturen inklusive molekularen Oberflächen und Eigenschaften zur Analyse vorlagen. Diese 7821 erfolgreich bearbeiteten Einträge entsprechen 82% der Datenbankeinträge, die Proteine enthalten, oder 76% aller Einträge der Proteindatenbank (Stand Oktober 1999).

5.1 Anwendung des vorgestellten Verfahrens auf die Proteindatenbank

5.1.1 Rechenzeitbedarf des Untersuchungsverfahrens

Pro Datenbankeintrag benötigt die Aufarbeitung der Protein-Struktur und die Berechnung der molekularen Oberfläche und Eigenschaften durchschnittlich 45 Minuten auf einem Prozessor eines SGI-Servers (MIPS R10000 196 MHz). Für alle Komplexe zusammen wurden insgesamt 5800 CPU-Stunden Rechenzeit benötigt. Tabelle 5.1 gibt einen Überblick über den Rechenzeitbedarf der einzelnen Schritte an vier repräsentativen Komplexen:

Tabelle 5.1: Rechenzeitbedarf der einzelnen Schritte an vier Beispielen:
1pam: 23143 Atome 1tup: 11543 Atome
2ptc: 4584 Atome 4pti: 1072 Atome.

Bearbeitungsschritt	1pam	1tup	2ptc	4pti
Analyse und Vorbereitung der Ausgangsdaten	3s	2s	1s	<1s
Ergänzung fehlender Strukturinformationen	4858s	3376s	682s	65s
Berechnung der molekularen Oberfläche	430s	281s	101s	28s
Berechnung und Projektion der Eigenschaften	6527s	2412s	852s	161s
Einteilung der Oberfläche in Teilbereiche	4600s	725s	245s	17s
Gesamtzeit	16418s	6796s	1880s	272s

5.1.2 Zusammensetzung und Aufbau der Proteinkomplexe

In den 7821 Datenbankeinträgen sind neben Strukturen einzelner Proteine auch Protein-Protein-, Protein-DNA- und Protein-Ligand-Komplexe vorhanden. Die Häufigkeiten der verschiedenen Komplextypen im Gesamtdatensatz sind in Tabelle 5.2 zusammengefaßt. Im Folgenden wird von einem Protein-Ligand-Komplex gesprochen, wenn der Komplexpartner des Proteins entweder ein nichtpeptidisches Molekül oder ein Oligopeptid mit weniger als neun Aminosäuren ist.

Tabelle 5.2: Anzahl der verschiedenen Komplextypen im Gesamtdatensatz.

Komplextyp:	Anzahl:
Kein Komplexbindungsbereich	3047
Komplex mit Protein-Protein-Bindung	2993
Komplex mit Protein-Ligand-Bindung	2716
Komplex mit Protein-DNA-Bindung	338
Komplex mit Protein-Protein- und Protein-DNA-Bindung	108
Komplex mit Protein-Protein- und Protein-Ligand-Bindung	1137
Komplex mit Protein-DNA- und Protein-Ligand-Bindung	68
Komplex mit Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindung	14

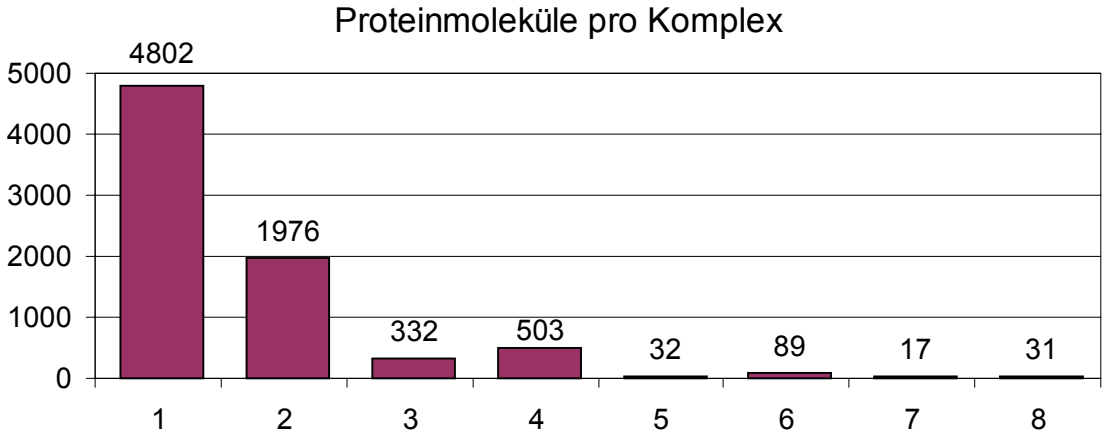


Diagramm 5.1: Aufteilung des Gesamtdatensatzes nach Anzahl der Proteinmoleküle pro Komplex.

Diagramm 5.1 gibt eine Zusammenstellung der Anzahl von Proteinmolekülen pro Komplex wieder. Den größten Anteil (61% = 4802 Strukturen) machen die Einzelproteinstrukturen aus. Ein Viertel der Gesamteinträge bzw. zwei Drittel der Protein-Protein-Komplexe (1976 Strukturen) sind Dimere. Die restlichen Protein-Protein-Komplexe

bestehen hauptsächlich aus drei, vier bzw. sechs Proteineinheiten. Die Verteilung zeigt Maxima bei gerader Anzahl von Proteinen pro Komplex (zwei, vier, sechs und acht Proteine). Protein-Protein-Komplexe können in zwei verschiedene Komplexarten eingeteilt werden. Zum einen gibt es Komplexe aus Proteinen mit identischer Aminosäuresequenz, solche Proteinkomplexe werden Oligomere genannt. Diese Quartärstruktur der Proteine ist häufig für die biologische Funktion von Proteinen wichtig (z.B.: Alkoholdehydrogenase-Dimer). Andererseits können Proteinkomplexe jedoch auch aus unterschiedlichen Polypeptidketten gebildet werden. Im Gesamtdatensatz A sind beide Komplexarten vertreten. Von den 2980 Protein-Protein-Komplexen sind 1908 Komplexe Oligomere. Der Hauptteil dieser Komplexe sind Dimere (1448 Strukturen). Tabelle 5.3 gibt einen Überblick über die Verteilung der unterschiedlichen Komplexe im Gesamtdatensatz A. 746 Protein-Protein-Komplexe bestehen nur aus unterschiedlichen Proteinen. Die restlichen 339 Komplexe werden sowohl aus unterschiedlichen Proteinen als auch aus Proteinen mit identischer Aminosäuresequenz gebildet.

Tabelle 5.3: Übersicht über die Häufigkeit der verschiedenen Proteinkomplexarten.

Polypeptidketten pro Komplex	2	3	4	5	6	7	8	Alle
Proteinkomplexe aus identischen Untereinheiten	1448	176	203	27	33	2	14	1908
Proteinkomplexe aus unterschiedlichen Untereinheiten	528	134	81	2	1	0	0	746
Gemischte Komplexe (sowohl identische als auch unterschiedliche Untereinheiten)	0	22	219	3	55	15	17	339

Insgesamt gibt es 38 verschiedene Protein-Protein-Komplextypen unter den 7821 bearbeiteten Proteinkomplexen. Die Komplextypen sowie ihre Häufigkeit im Gesamtdatensatz sind in Tabelle 5.4 aufgelistet und wie folgt zu interpretieren: Die Zahl gibt die Anzahl der Proteine mit identischer Aminosäuresequenz im Proteinkomplex an. Wenn Proteine mit unterschiedlicher Aminosäuresequenz einen Komplex bilden, sind sie durch ein „+“-Zeichen getrennt. So steht z.B. „1+1“ für einen Komplex aus zwei unterschiedlichen Proteinen, „3+1“ bedeutet, daß drei Proteine mit identischer Polypeptidkette und ein davon verschiedenes Protein einen Komplex bilden. Die Aufstellung zeigt nochmals, daß Oligomere im Datensatz stark vertreten sind. Neben diesen häufen sich auch Komplexe aus zwei oder mehr unterschiedlichen Oligomeren (z.B. 2+2, 2+2+2, 3+3, etc.). Beispiele für diese Komplextypen sind das Hämoglobin (2+2) oder die Aspartat-Transcarbamoylase (6+6).

Tabelle 5.4: Aufstellung der verschiedenen Protein-Protein-Komplextypen.

Anzahl Proteine pro Komplex	Anzahl Komplexe	Komplextyp	Anzahl Komplexe pro Typ
1	4802	1	4802
2	1976	1+1	528
		2	1448
3	332	1+1+1	134
		2+1	21
		3	177
4	503	1+1+1+1	81
		2+1+1	7
		2+2	210
		3+1	2
		4	203
5	32	1+1+1+1+1	2
		2+1+1+1	1
		2+2+1	2
		5	27
6	89	1+1+1+1+1+1	1
		2+2+1+1	1
		2+2+2	27
		3+2+1	1
		3+3	22
		4+2	2
		5+1	2
		6	33
7	17	2+1+1+1+1+1	1
		4+1+1+1	2
		4+3	3
		5+1+1	9
		7	2
8	31	2+2+2+2	1
		4+4	12
		6+2	1
		8	14
10	3	4+2+2+2	1
		10	2
11	1	11	1
12	9	4+4+4	1
		6+3+3	1
		6+6	6
		12	1

5.1.3 Einteilung der Daten in vier verschiedene Datensätze

Die Daten werden, wie in Kapitel 4.2.1, zur Analyse in vier Datensätze eingeteilt. Der Gesamtdatensatz A enthält alle verwendeten 7821 Strukturen. Der Datensatz B basiert auf der *PDB-Select*-Liste von Hobohm. Da diese Liste regelmäßig an den aktuellen Inhalt der *Protein Data Bank* angepaßt wird und zum anderen wie oben erwähnt nur 83% der Ausgangsstrukturen durch die hier entwickelte Methode bearbeitet werden konnten, ergibt sich eine Schnittmenge von 416 Strukturen, die in beiden Datensätzen enthalten sind und so den Datensatz B bilden. Der Antigen-Antikörper-Datensatz C sowie der Enzym-Inhibitor-Datensatz D wird durch eine Stichwortsuche erstellt. Es finden sich 226 Antigen-Antikörper- und 81 Enzym-Inhibitor-Komplexe. Tabelle 5.5 listet die Anzahl der Proteinstrukturen und Bindungsbereiche der Proteine in den vier Datensätzen auf. Protein-Protein- und Protein-DNA-Bindungsbereiche müssen eine Mindestgröße von 100 $Å^2$ haben und Protein-Ligand-Bindungsstellen müssen mindestens 20 $Å^2$ Proteinoberfläche besitzen.

Tabelle 5.5: Umfang der Datensätze A-D (A: Gesamtdatensatz, B: reduzierter Datensatz, C: Antigen-Antikörper-Komplexe, D: Enzym-Inhibitor-Komplexe).

Datensatz	Datenbank-einträge	Proteine	Protein-Protein-Bindungsbereiche	Protein-DNA-Bindungsbereiche	Protein-Ligand-Bindungsbereiche
A	7821	12931	5904	589	6810
B	416	661	304	47	272
C	226	606	419	2	212
D	81	201	133	0	37

5.2 Größe und Eigenschaften der Proteine und Bindungsbereiche

5.2.1 Größe, Oberfläche und Volumen der Proteine

Insgesamt sind in den 7821 Proteinkomplexen 12972 Proteinstrukturen enthalten. Im Mittel besteht jedes Protein aus 195 Aminosäuren bzw. 3024 Atomen. Die gemittelte molekulare Oberfläche pro Protein beträgt 8257 $Å^2$ und das mittlere Volumen 27079 $Å^3$. Tabelle 5.6 enthält die Mittelwerte für die verschiedenen Datensätze. Im Anhang sind die Größenangaben für die Proteinkomplexe in Tabelle 9.2 zusammengefaßt. Diagramm 5.2 zeigt die Größenverteilung der Proteine in den vier verschiedenen Datensätzen.

Tabelle 5.6: Mittlewerte und Standardabweichungen der Größenangaben von Proteinen unterteilt nach den Datensätzen A-D.

	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Atome pro Protein	3024 ± 2133	3639 ± 2146	2713 ± 1072	2319 ± 1764
Oberflächenatome	1621 ± 985	1903 ± 972	1533 ± 594	1215 ± 788
Aminosäuren pro Protein	195 ± 137	235 ± 138	178 ± 70	154 ± 115
Oberflächenaminosäuren	171 ± 114	203 ± 114	164 ± 64	133 ± 94
Molekülmasse [u]	21631 ± 15241	26031 ± 15340	19576 ± 7742	16672 ± 12608
Oberfläche [$Å^2$]	8257 ± 4763	9653 ± 4698	7935 ± 2995	6225 ± 3808
Volumen [$Å^3$]	27079 ± 19401	32691 ± 19595	24195 ± 9623	20788 ± 16027

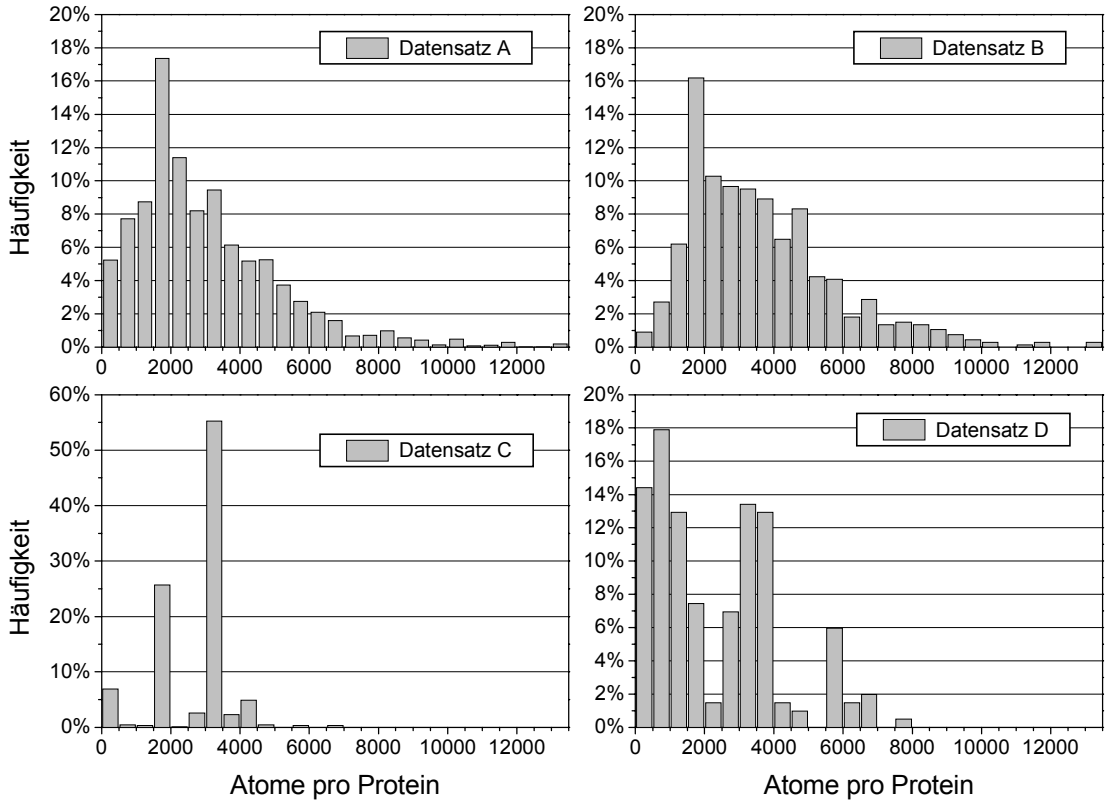


Diagramm 5.2: Größenverteilung der Proteine (Atome pro Protein) in den Datensätzen A-D. Im Diagramm für den Datensatz C wird eine andere Skala für die y-Achse (Häufigkeit) verwendet.

Die Größe der Proteine in der Proteindatenbank erstreckt sich über einen weiten Bereich [127]. Es gibt Proteinstrukturen bis zu einer Maximalgröße von 16246 Atomen bzw. 1025 Aminosäuren. Die Proteine im reduzierten Datensatz B sind im Mittel etwas größer als im Gesamtdatensatz A. Im Datensatz B sind weniger kleinere Proteine (siehe Diagramm 5.2) vorhanden. Die Proteine der Enzym-Inhibitor-Komplexe sind im Durchschnitt am

kleinsten. Die Inhibitoren in diesem Datensatz sind hauptsächlich kleine Proteine, wodurch der Mittelwert sinkt. Die Proteingrößenverteilung der Enzym-Inhibitor-Komplexe (Datensatz D) ist außerdem im Gegensatz zu den Datensätzen A und B nicht gleichmäßig. Es gibt drei Maxima in der Verteilung: Kleine Proteine mit weniger als 1000 Atomen (Inhibitoren) und Proteine mit ungefähr 3500 bzw. 6000 Atomen (Enzyme). Proteine mit anderen Größen sind nur wenige vorhanden. Im Antigen-Antikörper-Datensatz C sind hauptsächlich nur noch zwei verschiedene Größen von Proteinen vertreten (~2000 bzw. ~3000 Atome).

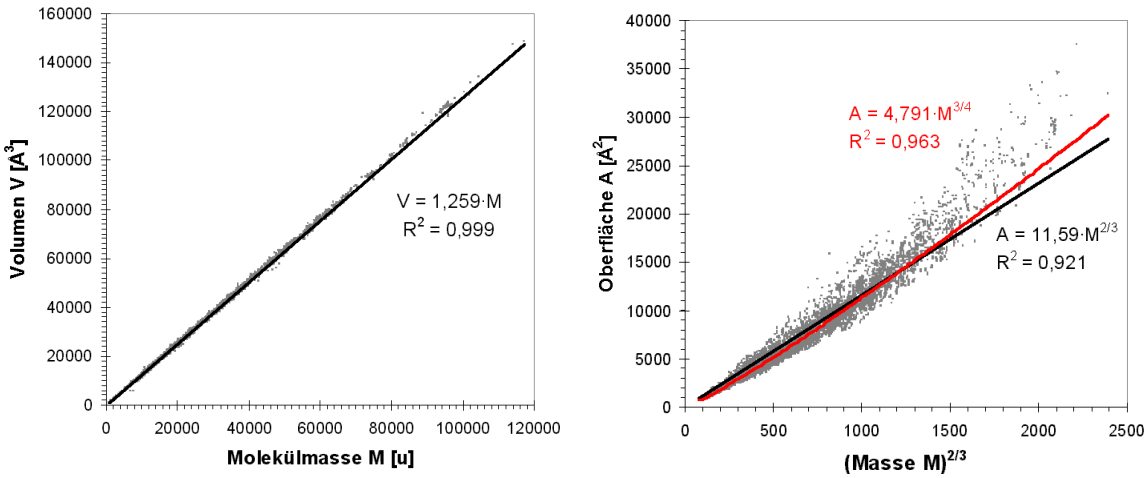


Diagramm 5.3: links) Verhältnis von Volumen zu Masse der untersuchten Proteine
rechts) Verhältnis von Oberfläche zu Masse ($M^{2/3}$) der Proteine.

Volumen und Oberfläche sind von der Molekülmasse der Proteine abhängig. In Diagramm 5.3 sind diese beiden Beziehungen dargestellt. Das Volumen der Proteine ist linear von der Masse der Proteine abhängig (siehe Diagramm 5.3 links). Die Dichte in den Proteinmolekülen ist unabhängig von der Molekülgröße und beträgt 2.1 g/cm^3. Auf der rechten Seite des Diagramms ist der Zusammenhang zwischen Oberfläche und Molekülmasse gezeigt. Für sphärische Körper steigt das Volumen mit der dritten Potenz, der Oberflächeninhalt aber nur mit der zweiten (d.h. $A \propto V^{2/3}$). Dementsprechend wird in dem Diagramm die Oberfläche gegen Masse$^{2/3}$ aufgetragen. Wenn die Proteine sphärische Körper sind, sollte eine lineare Abhängigkeit im Diagramm sichtbar sein (schwarze Trendlinie). Der Korrelationskoeffizient R^2 dieser Beziehung beträgt 0,92. Mit zunehmender Proteingröße steigt das Verhältnis Oberfläche und Masse ein wenig. Das Verhältnis wird besser durch eine exponentielle Regressionskurve ($A=4{,}791 \cdot M^{3/4}$, im Diagramm rot eingezeichnet) angenähert, deren Korrelationskoeffizient beträgt 0,96. Die

Zunahme der Oberfläche ist also proportional zu $V^{3/4}$. Dies ist ein Hinweis auf die fraktale Dimension von Moleküloberflächen [128-130]. Die Oberflächen von Molekülen sind weder glatt noch ideal sphärisch, sondern besitzen viele Spalten, Taschen und Erhebungen. Je größer die Proteine sind, desto mehr verschiebt sich das Verhältnis von Oberfläche und Masse zu einer linearen Abhängigkeit. Sehr große Proteine (Masse größer als 50000 a.u.) haben im Gegensatz zu den kleinen Proteinen meistens keine globuläre Form, sondern besitzen längliche Gestalt (z.B. Hexokinase Typ 1 – 1hkc). Dadurch erhöht sich das Oberfläche-Masse-Verhältnis.

5.2.2 Wasserstoffakzeptoren und Wasserstoffdonatoren

An der Proteinoberfläche befinden sich durchschnittlich etwa 400 Wasserstoffakzeptoren und Wasserstoffdonatoren (Tabelle 5.7). Über den Gesamtdatensatz gemittelt gibt es etwas mehr Akzeptoren als Donatoren an den Proteinoberflächen. Im Mittel ist jedes vierte bis fünfte Oberflächenatom ein Wasserstoffakzeptor oder -donator. Das entspricht einer durchschnittlichen Wasserstoffdonatoren- bzw. Wasserstoffakzeptorendichte von ca. 2,3 Akzeptoren/Donatoren pro 100 $Å^2$ molekulare Oberfläche. Die Dichten sind über alle Datensätze ungefähr gleich. Nur im Antigen-Antikörper-Datensatz sind sie ein wenig höher. Im Gegensatz zur Oberfläche gibt es mehr Donatoren als Akzeptoren in den Proteinen. Die Dichten der Donatoren an der Oberfläche (H-Donatoren pro Oberflächenatom) und im Gesamtprotein (H-Donatoren pro Atom) sind fast identisch.

Tabelle 5.7: Mittelwerte und Standardabweichungen der Anzahl und Dichte von Wasserstoffakzeptoren und -donatoren an der Proteinoberfläche und im Proteininneren. OFA ist die Abkürzung für „Oberflächenatom“.

	Datensatz A	Datensatz B	Datensatz C	Datensatz D
H-Donatoren an Oberfläche	187 ± 114	218 ± 112	192 ± 72	146 ± 97
H-Akzeptoren an Oberfläche	195 ± 123	230 ± 122	193 ± 76	149 ± 101
H-Donatoren im Protein	331 ± 234	394 ± 234	312 ± 130	257 ± 196
H-Akzeptoren im Protein	286 ± 205	343 ± 206	271 ± 114	222 ± 171
H-Donatorendichte [$Å^{-2}$]	0,0222 ± 0,0033	0,0224 ± 0,0026	0,0240 ± 0,0025	0,0225 ± 0,0033
H-Akzeptorendichte [$Å^{-2}$]	0,0228 ± 0,0030	0,0233 ± 0,0026	0,0237 ± 0,0027	0,0224 ± 0,0036
H-Donatoren pro OFA	11,5% ± 1,6%	11,5% ± 1,3%	12,6% ± 1,1%	11,9% ± 1,4%
H-Akzeptoren pro OFA	11,8% ± 1,3%	11,9% ± 1,1%	12,4% ± 1,0%	11,9% ± 1,3%
H-Donatoren pro Atom	11,4% ± 1,0%	11,3% ± 0,7%	11,8% ± 0,9%	11,5% ± 1,2%
H-Akzeptoren pro Atom	9,8% ± 0,7%	9,8% ± 0,5%	10,1% ± 0,7%	9,9% ± 0,7%

5.2.3 Größe und Eigenschaften der Bindungsbereiche

Die Größe der Bindungsbereiche der Proteine zu anderen Molekülen (Proteine, DNA-Moleküle und Liganden) sowie die Anzahl der Wasserstoffakzeptoren und -donatoren in diesen Bereichen sind in Tabelle 5.8 bis 5.10 aufgezeigt. Wenn ein Protein mit mehreren anderen Molekülen verbunden ist, werden alle Bindungsstellen einzeln unabhängig voneinander betrachtet.

Tabelle 5.8: Größe und Eigenschaften der Protein-Protein-Bindungsbereiche (Mittelwert und Standardabweichung). OFA ist die Abkürzung für „Oberflächenatom“.

	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Bindungsoberfläche [Å^2]	785 ± 669	883 ± 760	732 ± 467	604 ± 480
Anteil an Proteinoberfläche	11,5% ± 9,9%	11,1% ± 9,4%	11,3% ± 11,4%	14,1% ± 13,6%
Oberflächenatome	142,8 ± 125,4	161,4 ± 144,4	136,9 ± 89,5	110,7 ± 89,4
Aminosäuren	29,0 ± 22,2	32,5 ± 24,4	29,7 ± 18,7	23,3 ± 15,2
H-Donatoren	13,0 ± 11,8	14,2 ± 12,6	12,2 ± 7,4	10,7 ± 8,0
H-Akzeptoren	14,0 ± 12,6	16,1 ± 14,1	13,0 ± 8,3	11,5 ± 9,0
H-Donatorendichte [Å^{-2}]	0,0167 ± 0,0067	0,0163 ± 0,0063	0,0186 ± 0,0077	0,0179 ± 0,0067
H-Akzeptorendichte [Å^{-2}]	0,0178 ± 0,0048	0,0185 ± 0,0047	0,0182 ± 0,0053	0,0191 ± 0,0065
H-Donatoren pro OFA	9,5% ± 4,2%	9,3% ± 3,9%	10,5% ± 5,3%	10,0% ± 4,1%
H-Akzeptoren pro OFA	10,1% ± 2,9%	10,5% ± 3,1%	10,3% ± 4,3%	10,5% ± 3,2%
Wasserstoffbrücken	5,5 ± 5,9	6,3 ± 6,6	4,0 ± 2,7	5,6 ± 4,4
H-Brückendichte [Å^{-2}]	0,0069 ± 0,0050	0,0069 ± 0,0044	0,0063 ± 0,0047	0,0095 ± 0,0058
Disulfidbrücken	0,085 ± 0,362	0,048 ± 0,260	0,091 ± 0,295	0,135 ± 0,342

Die durchschnittliche Größe einer Protein-Protein-Bindungsregion auf der molekularen Proteinoberfläche ist 785 Å^2. Die Bindungsregionen sind im reduzierten Datensatz etwas größer (siehe Kapitel 5.2.1). Jedoch gibt es keine Beziehung zwischen Proteingröße und der Größe der Bindungsregionen (siehe auch Diagramm 5.4 und 5.5). In den Enzym-Inhibitor-Komplexen sind die Protein-Protein-Bindungsregionen aufgrund der kleinen proteinogenen Inhibitoren kleiner als in den anderen Datensätzen. Die 785 Å^2 Bindungsfläche (Gesamtdatensatz A) entsprechen etwa 11% der gesamten molekularen Oberfläche. In Enzym-Inhibitor-Komplexen ist dieser Anteil wegen der kleinen Inhibitoren größer (14%). An der Bildung der Protein-Protein-Bindungsstellen sind durchschnittlich 30 Aminosäuren beteiligt. Die Wasserstoffdonatoren- und Akzeptorendichte ist in den Protein-Protein-Bindungsbereichen (Tabelle 5.8) niedriger als an der Gesamtoberfläche (Tabelle 5.7). Wiederum sind über den Gesamtdatensatz gemittelt etwas mehr Akzeptoren

als Donatoren in den Proteinbindungsregionen vorhanden. Pro Protein-Protein-Bindung werden durchschnittlich fünf bis sechs Wasserstoffbrücken gebildet, welches einer Wasserstoffbrückendichte von etwa sieben Wasserstoffbrücken pro 1000 $Å^2$ Bindungsfläche entspricht. Von den 16 bis 18 Wasserstoffakzeptoren bzw. Wasserstoffdonatoren an der gleichen Fläche (1000 $Å^2$) werden ca. 15% zur Bildung von Wasserstoffbrückenbindungen verwendet. Disulfidbrücken sind sehr viel seltener. Über alle untersuchten Proteine betrachtet ist in jeder zwölften Protein-Protein-Bindungsstelle eine Disulfidbrücke vorhanden. Dieser Anteil schwankt jedoch sehr stark. In Enzym-Inhibitor-Komplexen sind fast doppelt so viel Disulfidbrücken vorhanden wie im Gesamtdatensatz A.

Tabelle 5.9: Größe und Eigenschaften von Protein-DNA-Bindungsbereichen.

	Datensatz A	Datensatz B
Bindungsoberfläche [$Å^2$]	480 ± 274	416 ± 279
Anteil an Proteinoberfläche	6,4% ± 5,0%	5,8% ± 4,3%
Oberflächenatome	83,9 ± 50,5	71,9 ± 53,4
Aminosäuren	23,7 ± 11,7	19,8 ± 10,9
H-Donatoren	19,0 ± 11,0	15,6 ± 10,5
H-Akzeptoren	7,0 ± 4,9	5,5 ± 4,7
H-Donatorendichte [$Å^{-2}$]	0,0400 ± 0,0086	0,0382 ± 0,0088
H-Akzeptorendichte [$Å^{-2}$]	0,0147 ± 0,0064	0,0127 ± 0,0060
H-Donatoren pro Oberflächenatom	23,4% ± 5,6%	22,9% ± 5,8%
H-Akzeptoren pro Oberflächenatom	8,4% ± 3,5%	7,5% ± 3,4%

Im Datensatz C und D sind fast keine oder überhaupt keine Protein-DNA-Komplexe vorhanden (siehe Tabelle 5.5), deswegen werden nur der Gesamtdatensatz A und der reduzierte Datensatz B zur Analyse der DNA-Bindungsregionen herangezogen. Die Protein-DNA-Bindungsbereiche sind etwas kleiner als die Protein-Protein-Bindungsregionen. Knapp 500 $Å^2$ Proteinoberfläche werden im Mittel für die Bindung eines DNA-Moleküls benötigt. Das entspricht etwa 6% der gesamten Proteinoberfläche. In den DNA-bindenden Oberflächenbereichen finden sich weniger Wasserstoffakzeptoren als an der Gesamtoberfläche und in den Protein-Protein-Bindungsbereichen. Dafür ist die Anzahl der Wasserstoffdonatoren stark erhöht. So sind fast zweimal so viele Donatoren im Protein-DNA-Bindungsbereich wie an der Gesamtoberfläche vorhanden. Fast ein Viertel (23,7%) der Oberflächenatome sind Wasserstoffdonatoren. Diese große Anzahl von Donatoren auf der Seite des Proteins und die negativ geladenen Sauerstoffatome (Akzeptoren) des Phosphatrückgrates der DNA-Moleküle können eine Vielzahl von Wasserstoffbrücken

ausbilden, wodurch eine starke Komplexbildung zustande kommt. Diese Besonderheit der Protein-DNA-Bindungsbereiche wird in späteren Kapiteln noch weiter vertieft (siehe Kapitel 5.4.3.3, 5.5.3 und 5.6).

Tabelle 5.10: Größe und Eigenschaften von Protein-Ligand-Bindungsbereichen. OFA ist die Abkürzung für „Oberflächenatom“.

	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Bindungsoberfläche [$Å^2$]	131 ± 107	154 ± 145	127 ± 141	194 ± 248
Anteil an Proteinoberfläche	1,5% ± 1,5%	1,6% ± 1,4%	1,4% ± 1,4%	3,4% ± 9,9%
Oberflächenatome	29,2 ± 24,6	34,5 ± 30,2	26,5 ± 28,9	43,5 ± 41,6
Aminosäuren	11,0 ± 7,0	12,3 ± 7,9	8,8 ± 7,5	14,2 ± 7,7
H-Donatoren	3,6 ± 3,7	4,4 ± 4,3	2,8 ± 3,2	4,3 ± 4,8
H-Akzeptoren	3,2 ± 3,0	3,8 ± 3,4	2,5 ± 3,4	4,9 ± 4,8
H-Donatorendichte [$Å^{-2}$]	0,0274 ± 0,0203	0,0298 ± 0,0205	0,0227 ± 0,0142	0,0273 ± 0,0217
H-Akzeptorendichte [$Å^{-2}$]	0,0238 ± 0,0146	0,0255 ± 0,0152	0,0182 ± 0,0130	0,0232 ± 0,0125
H-Donatoren pro OFA	13,4% ± 10,8%	13,8% ± 10,0%	11,9% ± 9,0%	13,3% ± 11,3%
H-Akzeptoren pro OFA	11,3% ± 7,3%	11,7% ± 7,4%	9,6% ± 8,7%	9,8% ± 4,7%

Protein-Ligand-Bindungsbereiche sind sehr viel kleiner als die Bindungsstellen von Proteinen zu anderen Protein- oder DNA-Molekülen. Die durchschnittliche Größe einer Ligandbindungsstelle ist etwa 130 $Å^2$, das entspricht ca. 1,5% der Gesamtoberfläche. Im Mittel sind 11 Aminosäuren an der Bildung des Bindungsbereiches beteiligt. Die Wasserstoffdonatoren- und Wasserstoffakzeptorendichte ist in den Ligandbindungsbereichen gegenüber der Gesamtoberfläche erhöht. Über alle untersuchten Proteine gemittelt ist die Wasserstoffdonatorendichte um etwa 25%, die Wasserstoffakzeptorendichte um 5% höher. Im reduzierten Datensatz sind beide Dichten sogar noch etwas höher, während sie im Antigen-Antikörper-Datensatz C niedriger sind. Die Ligandbindungsstellen im Datensatz C nehmen hier eine Sonderrolle in Bezug zur Wasserstoffakzeptoren- und Donatorendichte ein. In späteren Abschnitten der Arbeit wird gezeigt, daß die Ligandbindungsstellen im Datensatz C auch bei anderen molekularen Eigenschaften von den Werten in den anderen Datensätzen abweichen.

In den folgenden Diagrammen ist die Größenverteilung der Bindungsbereichoberfläche der Proteine (Protein-Protein, Protein-DNA- und Protein-Ligand-Bindungsbereiche) aufgetragen. In Diagramm 5.4 ist die absolute Größe und in Diagramm 5.5 die relative Größe (im Vergleich zur Gesamtoberfläche) dargestellt. Die Verteilung der Protein-Protein-Bindungsbereichgrößen ist sehr breit. Es gibt keine bevorzugte absolute und

relative Größe der Protein-Protein-Bindungsregionen an der molekularen Oberfläche. Bei den Protein-DNA-Bindungsbereichen gibt es ein Maximum bei ca. 400 Å^2 bzw. 4%. Jedoch gibt es auch einige DNA-Bindungsstellen, die sehr viel größer sind. Die Protein-Ligand-Bindungsbereiche sind meist sehr klein. Nur sehr wenige Bindungsflächen sind größer als 300 Å^2 bzw. 3%. Das Maximum liegt bei der vorgegebenen Minimalgröße von 20 Å^2. Es gibt keine eindeutige Abhängigkeit zwischen der Bindungsbereichgröße und der Gesamtoberfläche der Proteine (siehe Diagramm 5.5). Die Größe der Bindungsbereiche ist jeweils von der spezifischen Art und Funktion der Komplexbindung abhängig. Diese Zusammenhänge können mit den hier vorgestellten Methoden jedoch nicht erfaßt werden.

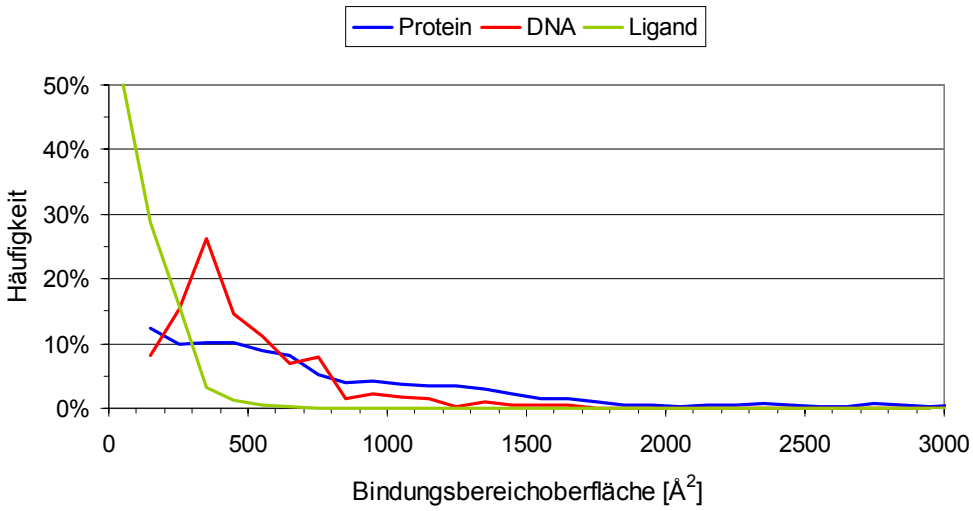


Diagramm 5.4: Größe der Bindungsbereiche (Protein-Protein, Protein-DNA und Protein-Ligand) an den molekularen Proteinoberflächen.

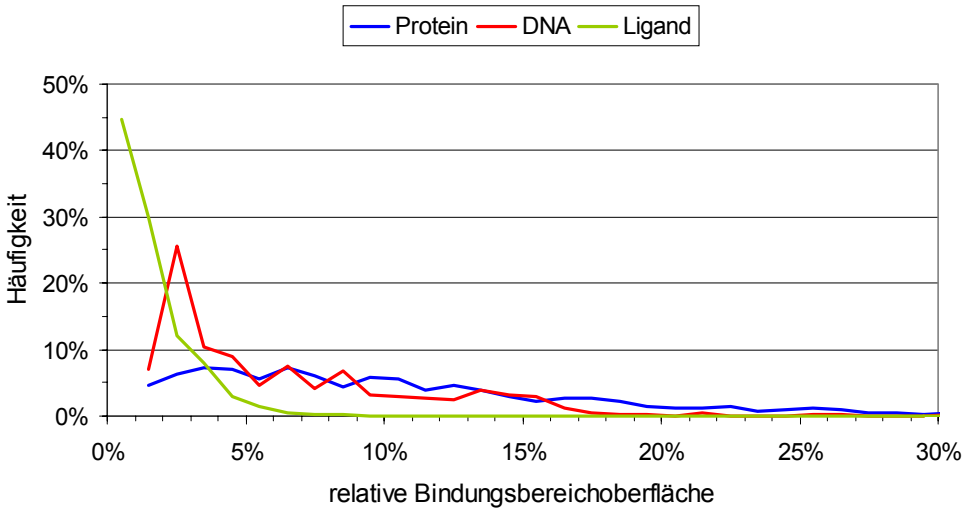


Diagramm 5.5: Relative Größe (in Bezug zur Gesamtproteinoberfläche) der Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereiche an den molekularen Proteinoberflächen.

5.3 Radiale Verteilungsfunktionen von H-Akzeptoren und H-Donatoren

In den Diagrammen 5.6 bis 5.10 sind die gemittelten radialen Verteilungsfunktionen g(r) für die vier verschiedenen Akzeptor/Donator-Paarabstände (Akzeptoren-Akzeptoren, Akzeptoren-Donatoren, Donatoren-Akzeptoren und Donatoren-Donatoren) in verschiedenen Bereichen der Proteinoberflächen dargestellt. Alle Verteilungsfunktionen sind auf die Wasserstoffakzeptoren- bzw. Wasserstoffdonatorendichte an der Proteinoberfläche (Tabelle 5.7) normiert, so daß sie gegen den Grenzwert 1,0 streben.

5.3.1 Verteilung der H-Akzeptoren und H-Donatoren an der Proteinoberfläche

Diagramm 5.6 zeigt die radialen Verteilungsfunktionen der Abstände zwischen Wasserstoffakzeptoren und Wasserstoffdonatoren an der gesamten Proteinoberfläche. Alle vier Funktionen besitzen deutliche Maxima und Minima. D.h. die Akzeptoren und Donatoren sind nicht gleichmäßig an der Außenseite der Proteine verteilt.

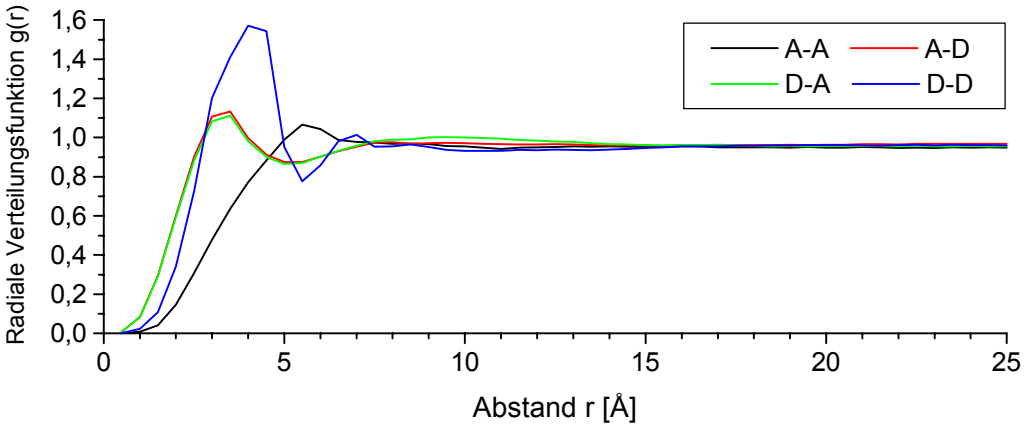


Diagramm 5.6: Radiale Verteilungsfunktionen (Gl. 3.11) der Abstände aller Wasserstoffakzeptoren und -donatoren an der Proteinoberfläche (Gesamtoberfläche):
A-A: Radiale Verteilungsfunktion von Akzeptor-Akzeptor-Abständen
A-D: Radiale Verteilungsfunktion von Akzeptor-Donator-Abständen
D-A: Radiale Verteilungsfunktion von Donator-Akzeptor-Abständen
D-D: Radiale Verteilungsfunktion von Donator-Donator-Abständen.

Die Verteilungsfunktion der Akzeptor-Akzeptor-Paarabstände steigt mit zunehmendem Abstand langsam an und durchläuft bei einem Abstand von ca. 5,5 Å ein leichtes Maximum. Dies entspricht dem entlang der molekularen Oberfläche gemessenen Abstand der beiden Sauerstoffatome in Carboxylgruppen, die z.B. in den Aminosäuren Glutamin- und Asparaginsäure vorkommen. Die Verteilungsfunktion für die Donator-Donator-

Abstände steigt etwas schneller an und besitzt zwei lokale Maxima und ein lokales Minimum. Das erste Maximum bei einem Donator-Donator-Abstand von 4 Å ist deutlich höher als das Maximum der Akzeptor-Akzeptor-Paare. Das zweite lokale Maximum bei dem Abstand 7 Å ist sehr viel niedriger und weniger ausgeprägt. Zwischen beiden Maxima durchläuft die Verteilungsfunktion das Minimum bei einem Donator-Donator-Abstand von 5,5 Å. Der Donator-Donator-Abstand von 4 Å (erstes Maximum) entspricht dem Abstand der Wasserstoffatome in der Ammoniumgruppe im Lysin, der NH_2-Gruppe im Arginin und der Amidgruppe im Glutamin bzw. Asparagin. Die Verläufe der radialen Verteilungsfunktionen der Akzeptor-Donator- und Donator-Akzeptor-Paarabstände sind wie erwartet fast identisch. Beide Funktionen steigen sehr schnell an und durchlaufen schon bei 3,5 Å das Maximum. Nach dem Maximum folgt bei einem Donator-Akzeptor-Abstand von ca. 5 Å das Minimum der radialen Verteilungsfunktionen.

Werden Akzeptor/Donator-Paare innerhalb einer Aminosäure nicht berücksichtigt (Kapitel 3.2.3.3), verändern sich die radialen Verteilungsfunktionen teilweise erheblich. Dies ist in Diagramm 5.7 dargestellt.

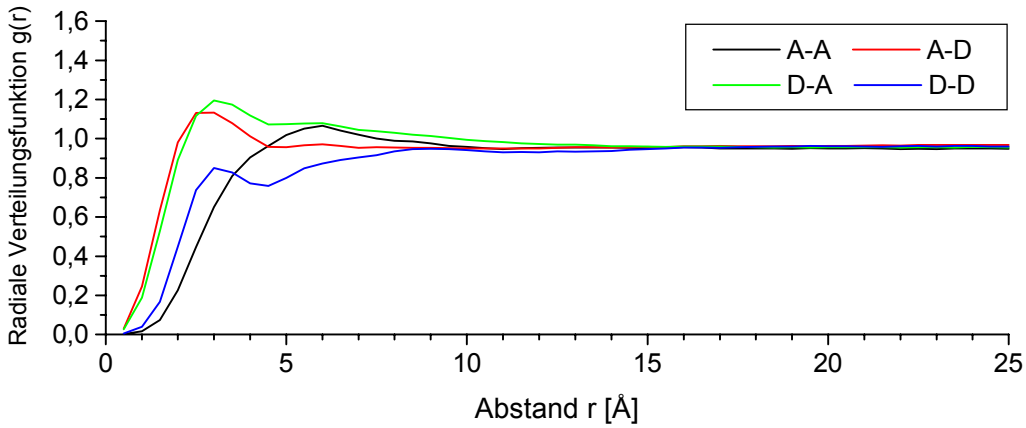


Diagramm 5.7: Radiale Verteilungsfunktionen der Abstände aller Wasserstoffakzeptoren und Wasserstoffdonatoren an der Proteinoberfläche (Gesamtoberfläche). Azeptor/Donator-Paare in der gleichen Aminosäure und deren Fläche werden nicht berücksichtigt (siehe Abbildung 3.4 rechts).

Die Verteilungsfunktion der Donator-Donator-Abstände unterscheidet sich im Vergleich zu der entsprechenden Funktion im Diagramm 5.6 deutlich. Das absolute Maximum bei 4 Å ist verschwunden. Es ist nur noch ein lokales Maximum bei ca. 3 Å vorhanden. Das Minimum ist ebenfalls nach links verschoben (Abstand r = 4,5 Å). Die deutlichen Maxima der radialen Verteilungsfunktionen der Donator-Akzeptor- und Akzeptor-Donator-

Abstände in Diagramm 5.6 sind auch in Diagramm 5.7 vorhanden (etwas zu kürzeren Abständen hin verschoben), jedoch sind die Minima der Verteilungsfunktionen bei 5 Å verschwunden. Weiterhin fällt auf, daß die Verteilungsfunktion der Donator-Akzeptor-Paare etwas höher liegt. Die Verteilung der Akzeptor-Akzeptor-Abstände ist hingegen fast unverändert, nur das Maximum ist etwas breiter.

Diese Veränderungen lassen den Schluß zu, daß Akzeptor/Donator-Paare, die von Atomen in der gleichen Aminosäure gebildet werden (z.B. von den beiden Wasserstoffatomen einer Amidgruppe), einen starken Einfluß auf den Verlauf der radialen Verteilungsfunktion im Bereich von kleinen Abständen r haben. D.h. die Maxima und Minima in den radialen Verteilungsfunktionen (Diagramm 5.6) resultieren zu einem großen Teil aus den weitgehend festgelegten Abständen zwischen den Atomen innerhalb einer einzelnen Aminosäure. Davon sind besonders die Donator-Donator-Paare betroffen. Im Gegensatz dazu deuten die Maxima der Donator-Akzeptor- und Akzeptor-Donator-Verteilungs-funktionen (Diagramm 5.7) darauf hin, daß es auch zwischen den Donatoren und Akzeptoren aus unterschiedlichen Aminosäuren bevorzugte Abstände gibt, sich also Verteilungsmuster an der molekularen Oberfläche bilden. Diese Muster werden unter anderem durch Wasserstoffbrücken zwischen den Aminosäuren hervorgerufen. Diese intramolekularen Wasserstoffbrücken bilden die Basis für die Sekundärstruktur der Proteine (α-Helix bzw. β–Faltblatt-Strukturen).

Die Ähnlichkeit der Akzeptor-Akzeptor-Funktionen in Diagramm 5.6 und 5.7 läßt vermuten, daß Akzeptor-Paare der selben Aminosäure (z.B. die beiden Sauerstoffatome der Carboxylgruppe) keinen Einfluß auf die radiale Verteilungsfunktion haben. Diese Ähnlichkeit ergibt sich jedoch aus dem unterschiedlichen Anteil der molekularen Oberfläche, der zur Berechnung der Verteilungsfunktionen verwendet wird (Gleichung 3.11). Um diesen Einfluß nachzuweisen, werden die Verteilungsfunktionen nochmals mit einer leicht veränderten Methode berechnet und in Diagramm 5.8 aufgetragen. Bei dieser Berechnung werden die Akzeptor/Donator-Paare innerhalb einer einzelnen Aminosäure nicht berücksichtigt, aber trotzdem die gesamte Ringfläche verwendet (im Unterschied zu Diagramm 5.7 bzw. Abbildung 3.4 rechts). Wenn die Akzeptor-Akzeptor-Paare innerhalb der selben Aminosäure wirklich keinen Einfluß auf den Verlauf der Verteilungsfunktion hätten, sollte diese unverändert bleiben. Die Akzeptor-Akzeptor-Verteilungsfunktion in Diagramm 5.8 hat allerdings an Stelle des Maximums nur noch eine Schulter im Bereich von 5,5 Å, wodurch der starke Einfluß der Akzeptor-Akzeptor-Paare innerhalb der selben Aminosäure auf den Verlauf der Verteilungsfunktion bestätigt wird.

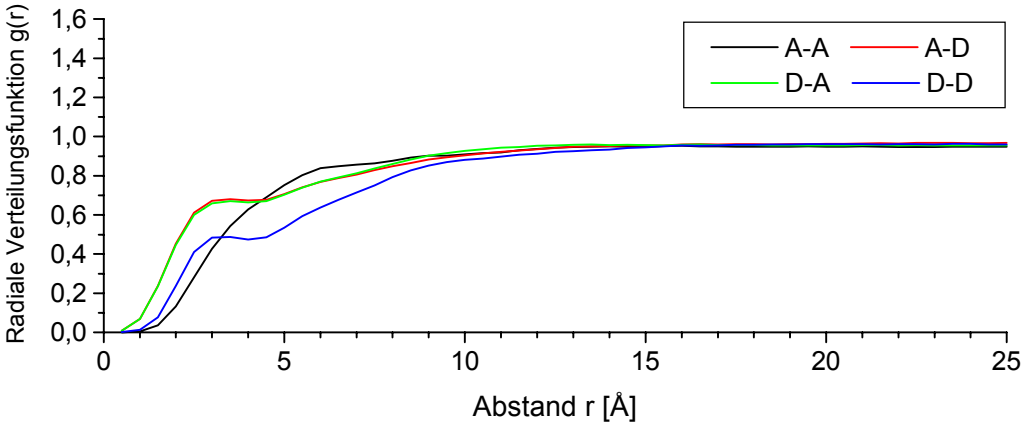


Diagramm 5.8: Radiale Verteilungsfunktionen der Abstände aller Wasserstoffakzeptoren und Wasserstoffdonatoren an der Proteinoberfläche. Akzeptor/Donator-Paare in der gleichen Aminosäure werden bei der Berechnung der Verteilungsfunktion (siehe Gleichung 3.11) nicht berücksichtigt. Im Gegensatz zu Diagramm 5.7 wird jedoch die gesamte Ringfläche in die Berechnung einbezogen.

5.3.2 Verteilung der H-Akzeptoren und H-Donatoren in Bindungsbereichen

In den nächsten Diagrammen wird untersucht, ob sich die Akzeptor/Donator-Muster in den Bindungsbereichen der Proteine verändern und somit die Komplexbildung beeinflussen. Die Verteilungsfunktionen der Akzeptor/Donator-Abstände in den Bindungsbereichen der molekularen Oberfläche zeigen teilweise starke Schwankungen. Dies betrifft besonders die Bindungsbereiche in Protein-DNA-Komplexen. In den untersuchten Komplexen ist nur eine geringe Anzahl von Protein-DNA-Komplexen vorhanden, so daß für die Berechnung der radialen Verteilungsfunktionen von Akzeptor/Donator-Paarabständen in diesen Bereichen sehr wenig Daten zur Verfügung stehen. Aus diesem Grund müssen etwaige lokale Maxima und Minima vorsichtig interpretiert werden.

Zuerst werden die radialen Verteilungsfunktionen mit Berücksichtigung von Akzeptor/Donator-Paaren innerhalb einer Aminosäure untersucht (Diagramm 5.9). In den vier Diagrammen werden jeweils die Verteilungsfunktion der Abstände aller Akzeptoren bzw. Donatoren an der Proteinoberfläche (d.h. den Akzeptor/Donator-Paaren an der Gesamtoberfläche) mit den Verteilungsfunktionen von Akzeptor/Donator-Paaren in den drei verschiedenen Typen von Bindungsbereichen verglichen.

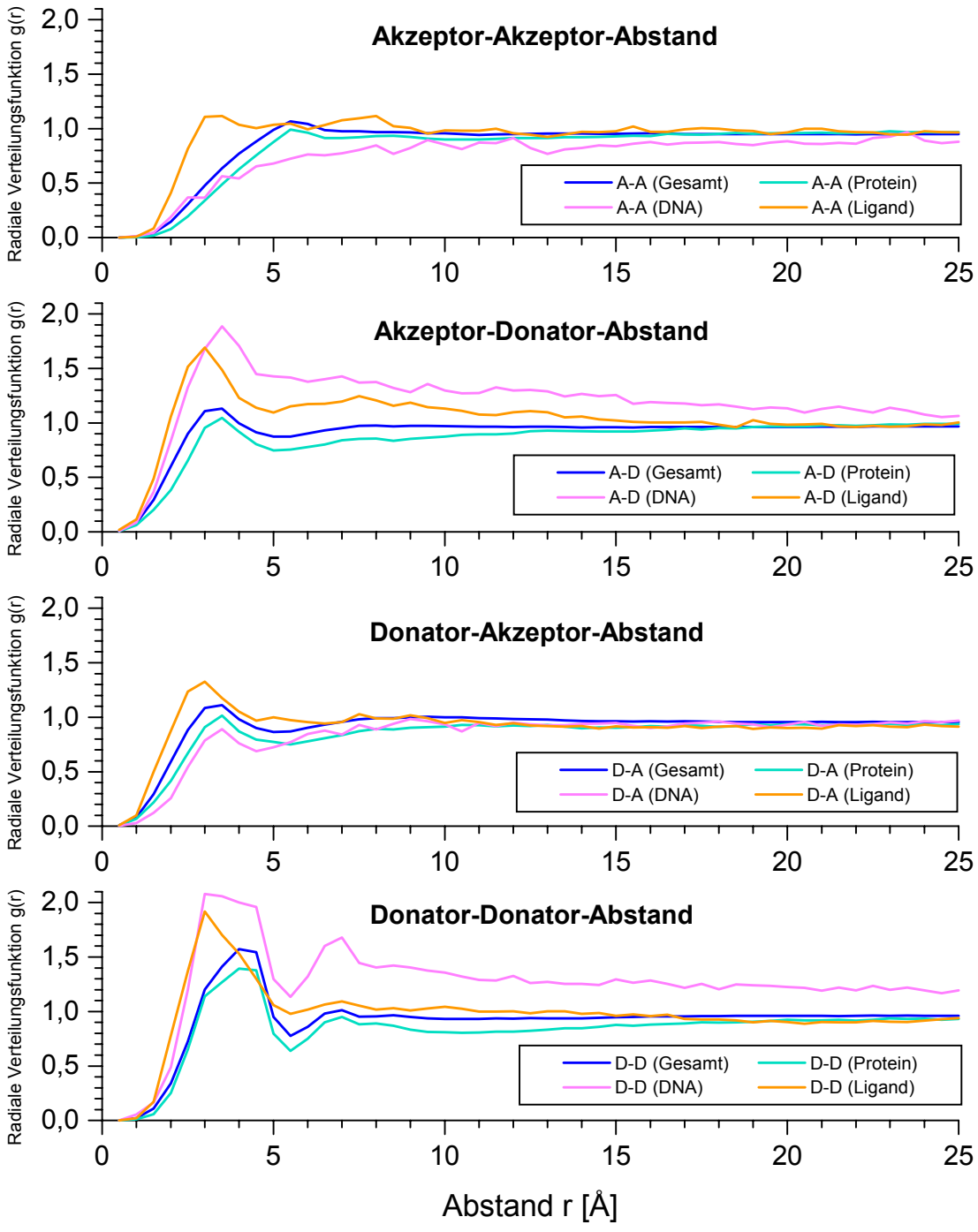


Diagramm 5.9: Vergleich der radialen Verteilungsfunktionen der Abstände von Akzeptor/Donator-Paaren an der gesamten Proteinoberfläche und in den Komplexbindungsbereichen (Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereiche).

In allen vier Diagrammen (Akzeptor-Akzeptor-, Akzeptor-Donator, Donator-Akzeptor- und Donator-Donator-Paare) sind die Verteilungsfunktionen der Akzeptor/Donator-Paare in den Protein-Protein-Bindungsbereichen sehr ähnlich zu den Funktionen der Paare an der Gesamtoberfläche, jedoch sind sie ein wenig nach unten verschoben. D.h. die Verteilung der Akzeptoren und Donatoren ist ähnlich wie an der Gesamtoberfläche, nur deren Anzahl ist geringer. Dies ist im Einklang mit den Werten in Tabelle 5.8. Sowohl die Wasserstoff-akzeptoren- als auch die Wasserstoffdonatorendichte ist in den Protein-Protein-Bindungs-regionen niedriger als über die gesamte Proteinoberfläche betrachtet.

Die Verteilungsfunktionen der Paare in den Protein-DNA- und Protein-Ligand-Bindungs-bereichen weichen hingegen stark von der Verteilung der Akzeptor/Donator-Paare an der gesamten Proteinoberfläche ab. Die radialen Verteilungsfunktionen von Paaren in den Ligandbindungsbereichen steigen schon bei sehr kurzen Abständen an. Dies ist besonders ausgeprägt in der Verteilungsfunktion der Akzeptor-Akzeptor-Paare. Dieses spezielle Verteilungsmuster in den Ligandbindungsbereichen ist auf die besondere Lage und Form derselben zurückzuführen. Liganden binden, wie in Kapitel 5.7.5 gezeigt werden wird, in konkaven, taschenförmigen Bereichen der molekularen Oberflächen. In diesen Bereichen sind die Abstände (entlang der Oberfläche gemessen) zwischen Atomen kürzer als in konvexen Oberflächenteilen (siehe auch Abbildung 5.1 in Kapitel 5.5.3). Im Gegensatz zu den Protein-Protein-Bindungsbereichen sind die radialen Verteilungsfunktionen der Akzeptor/Donator-Paare in den Bindungsregionen von Protein-Ligand-Komplexen etwas höher als an der Gesamtoberfläche. Dies ist besonders deutlich in den Funktionen der Akzeptor-Donator- und Donator-Donator-Paare sichtbar und auf die unterschiedlichen Wasserstoffakzeptor- und Donatordichten in den Ligandbindungsstellen zurückzuführen (siehe Tabelle 5.10).

Die größten Abweichungen zu den radialen Verteilungsfunktionen der Atompaare an der Gesamtoberfläche sind in den DNA-Bindungsbereichen zu finden. Die Lage von Minima und Maxima sind zwar sehr ähnlich zu den Funktionen der Paare an der gesamten Oberfläche, es gibt jedoch große Unterschiede in den Höhen der Verteilungsfunktionen: Die Verteilungsfunktionen der Akzeptor-Akzeptor-Paare und Donator-Akzeptor-Paare in den DNA-Bindungsregionen sind niedriger. Im Gegensatz dazu sind die Verteilungsfunktionen der Akzeptor-Donator- und Donator-Donator-Paare viel höher. Diese Beobachtungen können mit der hohen Anzahl von Wasserstoffdonatoren und der niedrigen Zahl von Wasserstoffakzeptoren in den DNA-Bindungsbereichen erklärt werden (siehe Tabelle 5.9).

Die radialen Verteilungsfunktionen der Abstände zwischen Wasserstoffakzeptoren und Wasserstoffdonatoren in den Bindungsbereichen der Proteinoberfläche ohne Berücksichtigung von Akzeptor/Donator-Paaren innerhalb der selben Aminosäure sind in Diagramm 5.10 zusammengestellt.

Die signifikantesten Unterschiede zu dem Diagramm 5.9 sind in den Protein-Protein-Bindungsbereichen vorhanden. Die Akzeptor-Donator-, Donator-Akzeptor- und Donator-Donator-Verteilungsfunktionen sind im Bereich unter 5 Å deutlich niedriger als die radialen Verteilungsfunktionen in Diagramm 5.9. Die Verteilungsfunktion der Akzeptor-Akzeptor-Paare hat sich wiederum kaum verändert.

Auch in den DNA-Bindungsbereichen der Proteinoberflächen zeigen sich zwei Unterschiede zu den in Diagramm 5.9 gezeigten Funktionen: Ähnlich wie in den Protein-Protein-Bindungsbereichen ist die Donator-Akzeptor-Verteilungsfunktion im Bereich bis 5 Å niedriger. Auch die Maxima der Donator-Donator-Verteilungsfunktion sind hier fast nicht mehr vorhanden. Das deutet darauf hin, daß diese Maxima durch Donator-Donator-Paare innerhalb einer Aminosäure hervorgerufen werden. Sehr großen Anteil an diesen Maxima haben die positiv geladenen Aminosäuren Lysin und Arginin, die sehr häufig in den DNA-Bindungsbereichen der molekularen Oberflächen vorkommen (siehe Kapitel 5.4.3.3). In diesen Aminosäuren gibt es vier bzw. fünf Wasserstoffdonatoren, die in einer weitgehend festgelegten geometrischen Anordnung zueinander stehen.

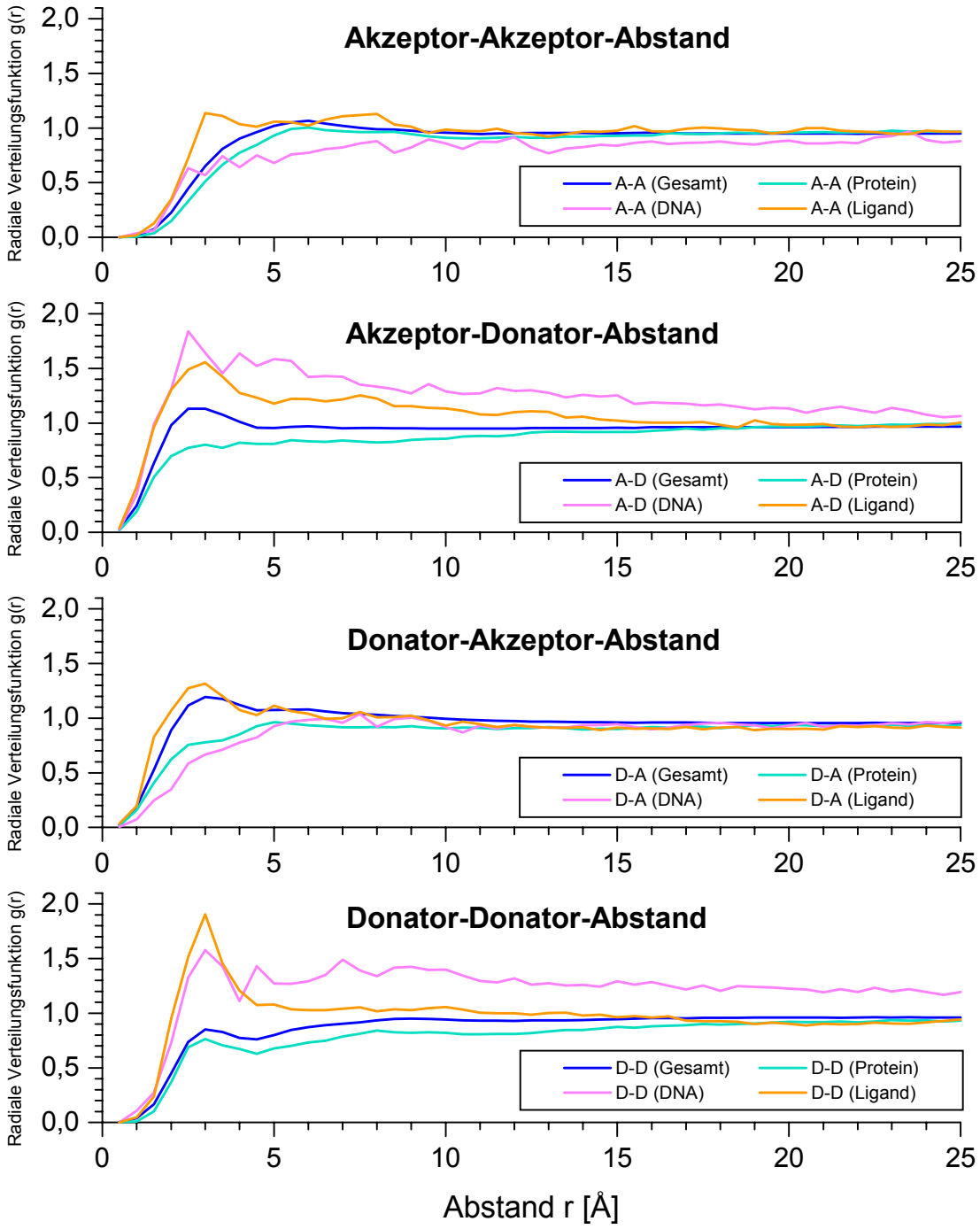


Diagramm 5.10: Vergleich der radialen Verteilungsfunktionen der Abstände von Akzeptor/Donator-Paaren an der gesamten Proteinoberfläche und in den Komplexbindungsbereichen (Protein-Protein, Protein-DNA und Protein-Ligand). Azeptor-Donator-Paare innerhalb der gleichen Aminosäure werden nicht berücksichtigt.

5.4 Verteilung der Aminosäuren im Protein und an der Oberfläche

5.4.1 Aminosäurenzusammensetzung der Proteine

Die 20 proteinogenen Aminosäuren sind in sehr unterschiedlicher Anzahl in Proteinen vorhanden. Diagramm 5.11 zeigt die mittleren relativen Häufigkeiten der proteinogenen Aminosäuren in den Proteinen des Gesamtdatensatzes im Vergleich mit den Untersuchungsergebnissen von McCaldon et al. [131]. Zwischen beiden Ergebnissen gibt es nur geringe Unterschiede der gemittelten Aminosäurenzusammensetzung. Die maximale Abweichung der mittleren relativen Häufigkeit beträgt 1,2% (Arginin). Die häufigste Aminosäure innerhalb der Proteine ist das hydrophobe Leucin. Cystein, Histidin, Methionin und Tryptophan sind hingegen in den Proteinstrukturen recht selten vorhanden.

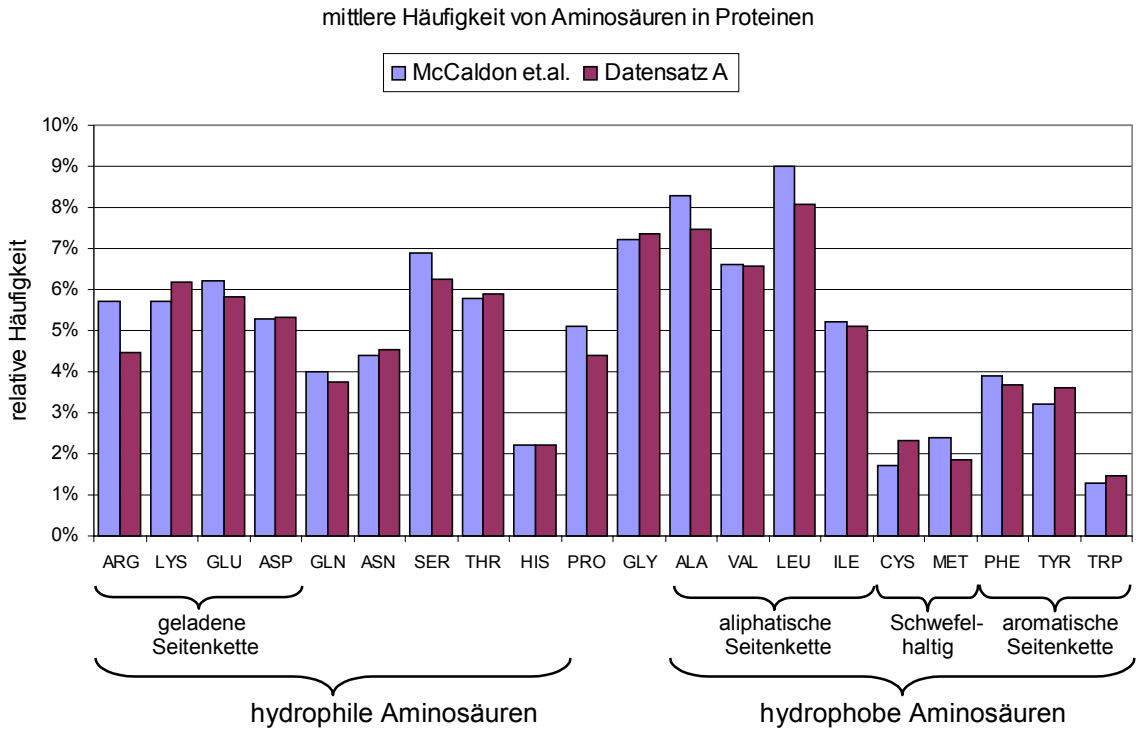


Diagramm 5.11: Aminosäurenzusammensetzung von Proteinen: Vergleich der mittleren Häufigkeiten der 20 natürlichen Aminosäuren innerhalb von Proteinen des Gesamtdatensatzes A mit den Ergebnissen von McCaldon [131].

In Tabelle 5.11 sind die gemittelten Häufigkeiten der einzelnen Aminosäuren mit den jeweiligen Standardabweichungen für die verschiedenen Datensätze A-D sowie die Ergebnisse von McCaldon et al. zusammengestellt.

Tabelle 5.11: Mittlere relative Häufigkeiten (inklusive Standardabweichungen) von Aminosäuren in Proteinen unterteilt nach den in dieser Arbeit verwendeten Datensätzen und den Ergebnissen von McCaldon et.al. [131]. Große Unterschiede der Häufigkeitswerte zwischen den Datensätzen sind grau hinterlegt. Referenzdatensatz ist dabei der Gesamtdatensatz A.

Aminosäure	[131]	Datensatz A	Datensatz B	Datensatz C	Datensatz D
ARG	5,7%	4,5% ± 3,1%	4,5% ± 2,5%	4,0% ± 3,0%	4,5% ± 4,0%
LYS	5,7%	6,2% ± 3,6%	5,9% ± 3,2%	5,5% ± 2,8%	5,0% ± 3,0%
GLU	6,2%	5,8% ± 3,3%	5,9% ± 2,9%	4,2% ± 2,7%	4,4% ± 4,6%
ASP	5,3%	5,3% ± 2,6%	5,6% ± 2,5%	4,6% ± 2,5%	4,6% ± 2,4%
GLN	4,0%	3,8% ± 2,3%	3,6% ± 2,0%	4,4% ± 2,0%	3,5% ± 2,3%
ASN	4,4%	4,5% ± 2,6%	4,7% ± 2,4%	3,8% ± 2,3%	5,3% ± 2,8%
SER	6,9%	6,2% ± 3,6%	5,9% ± 2,8%	11,8% ± 4,5%	7,7% ± 4,2%
THR	5,8%	5,9% ± 3,1%	5,8% ± 2,9%	8,8% ± 3,3%	6,3% ± 3,7%
HIS	2,2%	2,2% ± 1,9%	2,2% ± 1,8%	1,8% ± 1,5%	2,1% ± 2,6%
PRO	5,1%	4,4% ± 3,5%	4,4% ± 2,3%	5,3% ± 2,8%	5,4% ± 3,9%
GLY	7,2%	7,3% ± 3,5%	7,2% ± 3,1%	7,8% ± 4,0%	9,0% ± 3,5%
ALA	8,3%	7,5% ± 4,2%	7,6% ± 3,8%	5,9% ± 5,5%	8,3% ± 5,1%
VAL	6,6%	6,6% ± 3,1%	6,6% ± 2,8%	6,9% ± 2,7%	7,6% ± 4,7%
LEU	9,0%	8,1% ± 3,7%	7,9% ± 3,5%	7,4% ± 3,6%	6,6% ± 3,5%
ILE	5,2%	5,1% ± 3,0%	5,3% ± 2,6%	3,6% ± 2,5%	5,1% ± 2,8%
CYS	1,7%	2,3% ± 3,7%	1,9% ± 2,6%	2,1% ± 1,4%	4,4% ± 5,0%
MET	2,4%	1,8% ± 1,6%	2,0% ± 1,4%	1,5% ± 1,7%	1,2% ± 1,1%
PHE	3,9%	3,7% ± 2,2%	3,7% ± 1,9%	3,3% ± 2,2%	3,7% ± 2,9%
TYR	3,2%	3,6% ± 2,3%	3,6% ± 2,0%	5,0% ± 2,8%	3,9% ± 2,3%
TRP	1,3%	1,5% ± 1,3%	1,4% ± 1,2%	2,1% ± 1,2%	1,3% ± 1,2%

Die Datensätze A und B besitzen eine sehr ähnliche Verteilung der Aminosäuren. Die Differenz der mittleren relativen Häufigkeiten zwischen den beiden Datensätzen ist für keine Aminosäure größer als 1%. Das bedeutet, daß im Hinblick auf die Zusammensetzung der Proteine der reduzierte Datensatz B den Gesamtdatensatz A gut repräsentiert. Auch die Ergebnisse von McCaldon et al. stimmen sehr gut mit den in dieser Arbeit vorgestellten Daten überein. Diese Autoren fanden einen leicht höheren Anteil von Arginin. Die Abweichung resultiert aus der geringen Anzahl an untersuchten Strukturen. Die Ergebnisse von McCaldon et.al. basieren auf der Analyse von nur 1021 Proteinstrukturen. Die mittleren Zusammensetzungen der Proteine in den Datensätzen C (Antigen-Antikörper-Komplexe) und D (Enzym-Inhibitor-Komplexe) weichen teilweise stark von den Referenzdaten (Datensatz A) ab. So sind in den Antigen-Antikörper-Komplexen die

Aminosäuren Serin, Threonin und Tyrosin stärker vertreten. Sie sind gleichzeitig die beiden häufigsten Aminosäuren in den Proteinen des Antigen-Antikörper-Datensatzes C. Serin ist gegenüber der Aminosäurenzusammensetzung im Gesamtdatensatz A fast doppelt so oft vorhanden. Proteine in den Enzym-Inhibitor-Komplexen weisen besonders viel Cystein auf.

5.4.2 Aminosäurenzusammensetzung der Proteinoberfläche

Viele Untersuchungen zeigen, daß Proteine einen hydrophoben Kern und eine hydrophile Außenseite besitzen [z.B. 33]. Das Zusammenlagern von hydrophoben Seitenketten im Proteininneren ist eine Triebkraft für die Faltung der Proteinstruktur [36]. Dieses sollte sich auch in der Aminosäurenverteilung widerspiegeln.

Tabelle 5.12 faßt die Aminosäurenverteilung an der Proteinoberfläche in den verschiedenen Datensätzen zusammen, und Diagramm 5.12 präsentiert die Unterschiede zwischen den einzelnen Datensätzen. Es fällt wiederum auf, daß die Werte für die Proteine der Antigen-Antikörper-Komplexe teilweise sehr stark von den anderen Datensätzen abweichen. Serin und Threonin sind, wie auch bei der Verteilung über das gesamte Protein gezeigt wurde, sehr viel häufiger vorhanden.

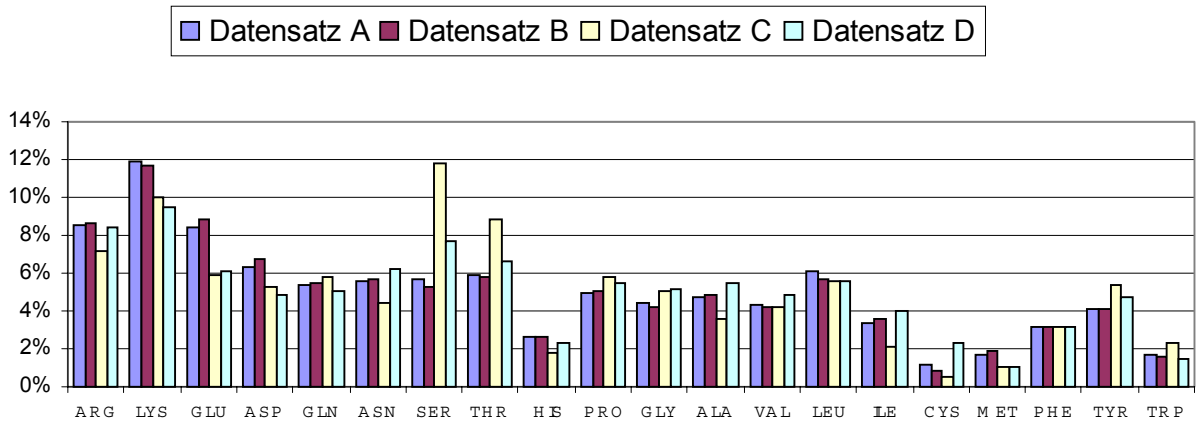


Diagramm 5.12: Aminosäurenzusammensetzung an Proteinoberflächen - Vergleich der Zusammensetzung der Datensätze A-D.

Tabelle 5.12: Mittelwerte und Standardabweichungen der Anteile der 20 proteinogenen Aminosäuren an der Gesamtoberfläche der Proteine in den vier verschiedenen Datensätzen A-D. Starke Unterschiede zwischen den Datensätzen sind grau hervorgehoben. Referenzdatensatz ist dabei der Gesamtdatensatz A.

Aminosäure	Datensatz A	Datensatz B	Datensatz C	Datensatz D
ARG	8,5% ± 5,5%	8,6% ± 4,5%	7,1% ± 5,3%	8,4% ± 7,4%
LYS	11,9% ± 6,2%	11,7% ± 5,5%	10,0% ± 4,4%	9,5% ± 5,6%
GLU	8,4% ± 4,3%	8,8% ± 4,1%	5,9% ± 3,6%	6,1% ± 5,6%
ASP	6,3% ± 3,1%	6,7% ± 3,0%	5,3% ± 2,6%	4,9% ± 2,7%
GLN	5,4% ± 2,9%	5,4% ± 2,8%	5,8% ± 2,7%	5,1% ± 3,3%
ASN	5,6% ± 3,3%	5,7% ± 3,0%	4,4% ± 2,9%	6,2% ± 3,9%
SER	5,6% ± 3,8%	5,2% ± 2,9%	11,8% ± 5,1%	7,6% ± 5,1%
THR	5,9% ± 3,2%	5,8% ± 3,0%	8,9% ± 3,6%	6,6% ± 4,1%
HIS	2,7% ± 2,4%	2,6% ± 2,4%	1,8% ± 1,8%	2,3% ± 3,0%
PRO	5,0% ± 3,5%	5,0% ± 2,5%	5,8% ± 2,5%	5,4% ± 3,5%
GLY	4,4% ± 2,1%	4,2% ± 1,8%	5,1% ± 2,6%	5,2% ± 2,4%
ALA	4,8% ± 3,2%	4,9% ± 2,9%	3,6% ± 4,7%	5,5% ± 4,2%
VAL	4,3% ± 2,5%	4,2% ± 2,0%	4,3% ± 2,4%	4,9% ± 4,2%
LEU	6,1% ± 3,5%	5,7% ± 3,1%	5,5% ± 4,4%	5,6% ± 3,8%
ILE	3,4% ± 2,8%	3,6% ± 2,2%	2,1% ± 2,7%	4,0% ± 3,3%
CYS	1,1% ± 2,2%	0,9% ± 1,5%	0,5% ± 1,1%	2,4% ± 3,4%
MET	1,7% ± 1,9%	1,9% ± 1,6%	1,1% ± 2,2%	1,0% ± 1,5%
PHE	3,2% ± 2,5%	3,2% ± 1,9%	3,2% ± 2,7%	3,2% ± 2,6%
TYR	4,1% ± 3,0%	4,1% ± 2,5%	5,4% ± 4,0%	4,7% ± 2,9%
TRP	1,6% ± 1,8%	1,6% ± 1,5%	2,3% ± 1,8%	1,5% ± 1,8%

Zur besseren Verdeutlichung des Unterschiedes zwischen Proteinkern und Proteinoberfläche ist in Diagramm 5.13 die Differenz zwischen der Aminosäurenzusammensetzung an der Proteinoberfläche und der Zusammensetzung im gesamten Protein dargestellt. Wie erwartet sind an der Molekülоberfläche die hydrophilen Aminosäuren, insbesondere die geladenen Aminosäuren (Lysin, Arginin, Glutaminsäure und Asparaginsäure), häufiger als im Inneren des Proteins anzutreffen. Lysin hat den größten Oberflächenanteil aller Aminosäuren. Lysin und Arginin sind im Vergleich zu ihrer Verteilung im gesamten Protein an der Proteinoberfläche fast doppelt so häufig vorhanden. Die hydrophoben Aminosäuren (Alanin, Valin, Leucin und Isoleucin) sind an der Oberfläche weniger vorhanden. Cystein ist ebenfalls an der Molekülоberfläche weniger häufig vertreten als im Proteininneren. Die restlichen Aminosäuren (Serin, Threonin,

Histidin, Methionin, Phenylalanin, Tyrosin, Prolin und Tryptophan) sind an der Oberfläche mit fast der gleichen Häufigkeit wie im Proteininneren vertreten. Interessant ist hierbei, daß dieses im Antigen-Antikörper-Datensatz C auch für Serin und Threonin gilt, die dort sowohl im Proteininneren als auch an der Proteinoberfläche die häufigsten Aminosäuren sind.

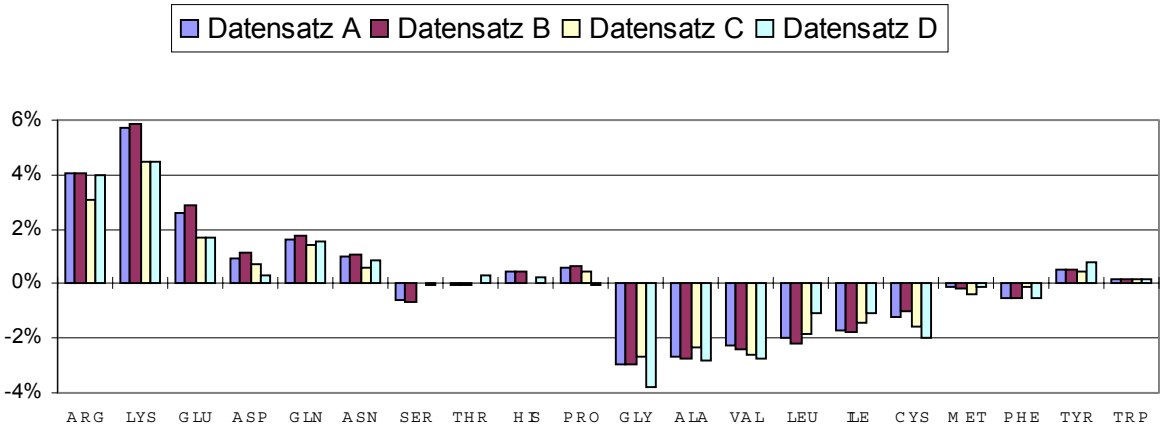


Diagramm 5.13: Aminosäurenzusammensetzung von Proteinen bzw. Proteinoberflächen: Differenz (a_{rel} - n_{rel}) zwischen der relativen Häufigkeit der proteinogenen Aminosäuren an der Proteinoberfläche und im gesamten Protein (siehe Kapitel 4.2.5)

5.4.3 Aminosäurenzusammensetzung von Bindungsbereichen

Untersuchungen der Aminosäurenzusammensetzung von Proteinen, Proteinoberflächen und deren Bindungsbereichen zeigen, daß die Bindungsbereiche in Protein-Protein-Komplexen mehr dem Proteininneren als der Oberfläche ähneln [17,19,20,26,33]. Diese Untersuchungen wurden bisher aber nur an kleineren Datensätzen vorgenommen. Deswegen werden in dieser Arbeit die Analysen den gesamten, in der *Protein Data Bank* zur Verfügung stehenden Strukturinformationen (Gesamtdatensatz A) erneut durchgeführt. In den folgenden Diagrammen 5.14 bis 5.16 sind die Unterschiede (Differenz) zwischen der Aminosäurenzusammensetzung in den Bindungsbereichen der Proteinoberfläche und der Zusammensetzung an der gesamten Oberfläche dargestellt. Dabei wird die Art der Bindungsregion (Protein-Protein, Protein-Ligand oder Protein-DNA) unterschieden. Eine vollständige Zusammenstellung der gemittelten relativen Aminosäurehäufigkeiten und den dazugehörigen Standardabweichungen ist in den Tabellen 9.3 bis 9.5 im Anhang gegeben.

5.4.3.1 Protein-Protein-Bindungsbereiche

Die Untersuchung aller Proteine des Gesamtdatensatzes (siehe Diagramm 5.14) bestätigt die Ergebnisse der Untersuchungen an kleineren Datensätzen [17,19,20,26,33].

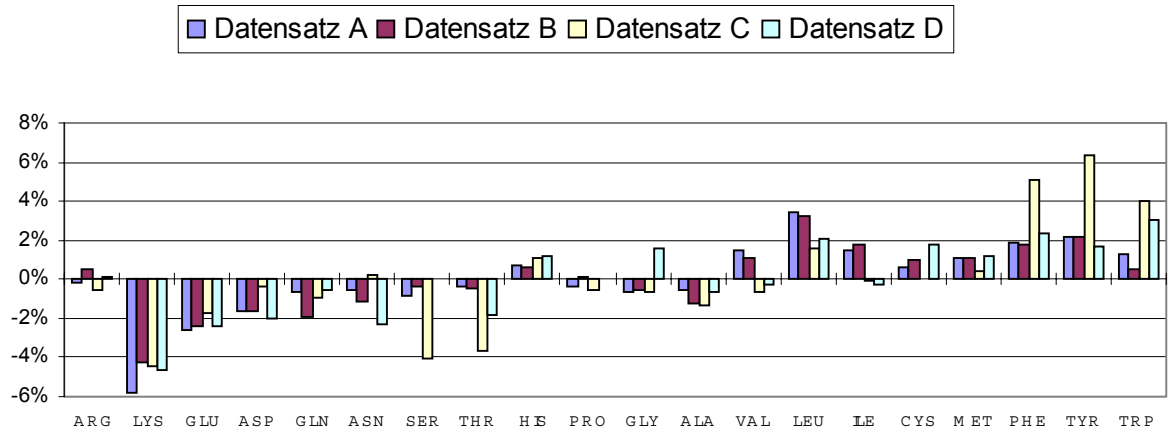


Diagramm 5.14: Differenz ($a_{rel,Protein\text{-}Protein} - a_{rel,gesamt}$) der Aminosäurenzusammensetzung an der Oberfläche von Protein-Protein-Bindungsbereichen und der gesamten Proteinoberfläche.

Die Protein-Protein-Bindungsstellen enthalten im Vergleich zur nichtbindenden Proteinoberfläche überwiegend mehr hydrophobe Aminosäuren (Leucin, Methionin, Phenylalanin, Tyrosin und Tryptophan). Die hydrophilen Aminosäuren (Lysin, Glutaminsäure, Asparaginsäure, Serin und Threonin) sind dagegen weniger oft vorhanden. Eine Sonderstellung nimmt das zu den hydrophilen Aminosäuren zählende Arginin ein. Der Anteil von Arginin an der Oberfläche ist jedoch im Gegensatz zu den anderen hydrophilen Aminosäuren im Protein-Protein-Bindungsbereich ähnlich wie an der restlichen Oberfläche des Proteins. Zusätzlich ist Arginin zusammen mit Leucin auch die Aminosäure mit dem größten Oberflächenanteil im Bindungsbereich, dicht gefolgt von Tyrosin (siehe Tabelle 9.3 im Anhang). Diese hohe Häufigkeit von Arginin in den Protein-Protein-Bindungsbereichen könnte als ein Hinweis auf die Anwesenheit von vielen Salzbrücken zwischen den Proteinmolekülen gewertet werden. Jedoch sind die für eine Salzbrücke nötigen negativ geladenen Aminosäuren Glutamin- und Asparaginsäure in den Bindungsbereichen deutlich reduziert. Arginin kann aber auch mit den aromatischen Seitenketten anderer Aminosäuren (Phenylalanin, Tyrosin und Tryptophan) Kontakte zwischen den Komplexpartnern ausbilden [27,29]. Diese Aminosäuren sind in den Bindungsbereichen deutlich häufiger als in den restlichen Bereichen der molekularen Oberfläche vorhanden, was für die Ausbildung solcher Kontakte spricht. Die leichte Erhöhung der Häufigkeit von Cystein

in den Bindungsbereichen der Protein-Protein-Komplexe deutet auf mögliche Disulfidbrücken zwischen den Komplexpartnern hin. Dies wird in dem späteren Kapitel 5.6 noch genauer untersucht werden. Die Aminosäurenzusammensetzung der Protein-Protein-Bindungsbereiche ist in dem Gesamtdatensatz A und dem reduzierten Datensatz B sehr ähnlich. Die Proteine des Antigen-Antikörper-Datensatzes weichen jedoch stark ab. Starke Unterschiede sind hier wieder bei Serin und Threonin zu beobachten. Sie sind im Protein-Protein-Bindungsbereich seltener zu finden. Der stark erhöhte Anteil (gegenüber dem Gesamtdatensatz) der beiden Aminosäuren an der Gesamtoberfläche bzw. innerhalb der Proteine des Antigen-Antikörper-Datensatzes ist im Bindungsbereich nicht sichtbar. Die beiden Aminosäuren sind also im Protein-Protein-Bindungsbereich abgereichert. Die Aminosäuren Phenylalanin, Tyrosin und Tryptophan (aromatischen Seitenketten) sind hingegen in den Protein-Protein-Bindungsbereichen der Antigen-Antikörper-Komplexe zahlreicher vorhanden. Diese Anhäufung wurde auch schon von anderen Autoren berichtet und erklärt [27]. Die aromatischen Ringsysteme können Wechselwirkungen zwischen den Proteinen ausbilden und tragen so zu Komplexstabilität bei. Die Zusammensetzung der Bindungsbereiche von Enzym-Inhibitor-Komplexen ist ähnlich zu den Komplexen in Datensatz A und B. Kleinere Unterschiede werden bei den Aminosäuren Glycin und Tryptophan beobachtet. Sie sind etwas öfter in den Protein-Protein-Bindungsbereichen der Enzym-Inhibitor-Komplexe vorhanden.

5.4.3.2 Protein-Ligand-Bindungsbereiche

In Tabelle 9.4 im Anhang sind die Anteile der Aminosäuren an der Oberfläche von Ligandbindungsbereichen für alle vier Datensätze aufgelistet. Die Werte für die Bindungsbereiche in den Datensätzen C und D weisen teilweise sehr hohe Standardabweichungen und damit eine geringe statistische Signifikanz auf. Grund hierfür ist die geringe Anzahl von Protein-Ligand-Komplexen in diesen Datensätzen (siehe Tabelle 5.5). Die Analyse der Aminosäurenzusammensetzung der Ligandbindungsoberfläche wird deshalb auf die Datensätze A und B beschränkt (Diagramm 5.15). Zwischen diesen beiden Datensätzen gibt es wie bei den Protein-Protein-Bindungsbereichen nur geringe Unterschiede. Die Zusammensetzung der Ligandbindungsoberflächen ähnelt den Protein-Protein-Bindungsflächen. Die hydrophilen bzw. geladenen Aminosäuren haben im Ligandbindungsbereich einen niedrigeren Oberflächenanteil, während die hydrophoben Aminosäuren öfters vorkommen. Prolin ist deutlich weniger vorhanden. Eine Ausnahme ist erneut Arginin, dessen Oberflächenanteil mit dem Wert an der gesamten Proteinoberfläche übereinstimmt

und damit die häufigste Aminosäure in den Ligandbindungsbereichen ist. Einen sehr hohen Anteil besitzen auch Leucin und die aromatischen Aminosäuren Tyrosin, Phenylalanin und Tryptophan, die gegenüber der Gesamtoberfläche deutlich häufiger vorhanden sind und somit zur Ligandbindung mit ihren aromatischen Ringsystemen beitragen.

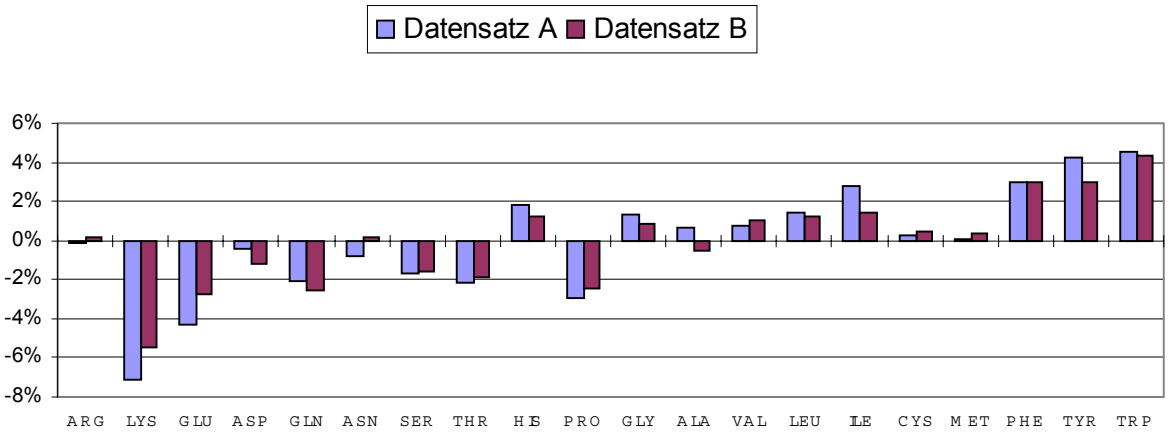


Diagramm 5.15: Differenz der Aminosäurenzusammensetzung an der Oberfläche von Protein-Ligand-Bindungsbereichen und der gesamten Proteinoberfläche.

5.4.3.3 Protein-DNA-Bindungsbereiche

Protein-DNA-Komplexe kommen in den untersuchten Daten im Gegensatz zu Protein-Protein- und Protein-Ligand-Komplexen nur selten vor (siehe Tabelle 5.5). Im Enzym-Inhibitor-Datensatz sind keine und im Antigen-Antikörper-Datensatz nur drei Protein-DNA-Komplexe vorhanden. Die Auswertung der Protein-DNA-Bindungsbereiche wird deshalb wie bei den Ligandkomplexen auf die Datensätze A und B beschränkt (Tabelle 9.5 im Anhang). Auch zwischen diesen Datensätzen sind teilweise große Unterschiede in den Oberflächenanteilen der Aminosäuren vorhanden, die ebenfalls auf die geringe Anzahl von Protein-DNA-Bindungsbereichen in den Datensätzen zurückzuführen sind. Diagramm 5.16 zeigt die Differenz der Oberflächenanteile im DNA-Bindungsbereich und an der gesamten Proteinoberfläche. Der Anteil der positiv geladenen Aminosäure Arginin an der Oberfläche der DNA-Bindungsbereiche ist stark erhöht. Im Gesamtdatensatz ist der Anteil von Lysin ebenfalls leicht erhöht. Die negativ geladenen Aminosäuren Glutaminsäure und Asparaginsäure sind deutlich weniger vorhanden. Bei den restlichen Aminosäuren ist kein klarer Trend sichtbar. Valin, Leucin und Prolin haben einen niedrigeren Oberflächenanteil als an der gesamten Proteinoberfläche, Serin und Threonin sind hingegen häufiger im DNA-Bindungsbereich anzutreffen. Die positiv geladenen Aminosäuren Lysin und Arginin

haben den höchsten Oberflächenanteil aller Aminosäuren im Protein-DNA-Bindungsbereich (Tabelle 9.5 im Anhang). Diese Verteilung der positiven und negativen Aminosäuren zeigt deutlich den Charakter von Protein-DNA-Bindungen, der durch attraktive Coulomb-Wechselwirkungen zwischen der elektronegativ polarisierten DNA-Oberfläche (hervorgerufen durch die negativ geladenen Phosphatgruppen) und der überwiegend elektropositiven Proteinaußenseite geprägt ist [37,132].

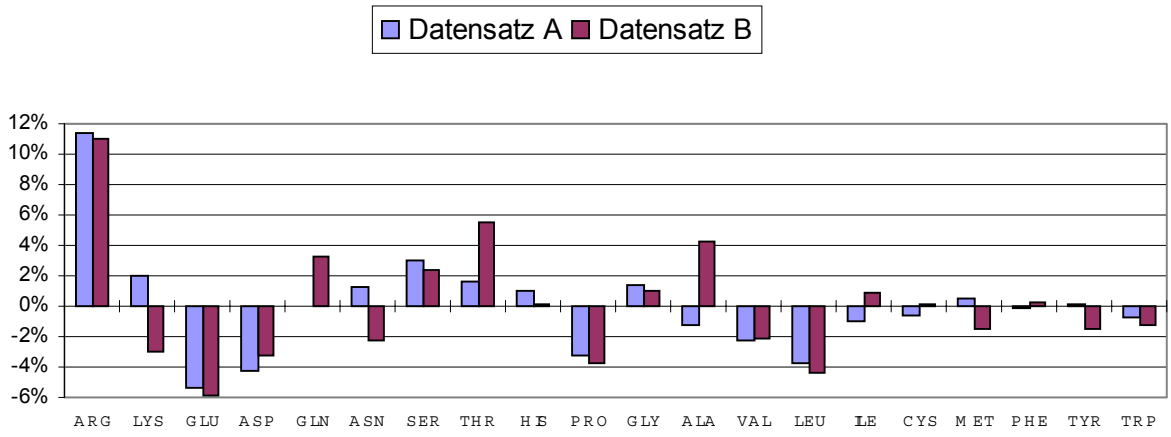


Diagramm 5.16: Differenz der Aminosäurenzusammensetzung an der Oberfläche von Protein-DNA-Bindungsbereichen und der gesamten Proteinoberfläche.

5.5 Molekulare Oberfläche einzelner Aminosäuren

5.5.1 Oberfläche von einzelnen Aminosäuren in Tripeptiden

Neben dem Anteil der Aminosäuren an der molekularen Oberfläche ist auch die Fläche, die den einzelnen Aminosäuren an der Proteinoberfläche zugeordnet wird, interessant. Entsprechend der Beschreibung in Kapitel 4.2.5 werden die mittleren Flächeninhalte der Aminosäuren an den Proteinoberflächen berechnet und mit den Werten der Tripeptide Glycin-X-Glycin (X = 20 proteinogene Aminosäuren) verglichen. In Tabelle 9.6 und 9.7 sind die Flächeninhalte der Aminosäuren bzw. von deren Seitenketten für die Tripeptide und Proteinoberflächen zusammengefaßt. Die Werte für die Proteinoberfläche sind nach der Gesamtoberfläche und den verschiedenen Bindungsbereichen (Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereich) unterschieden.

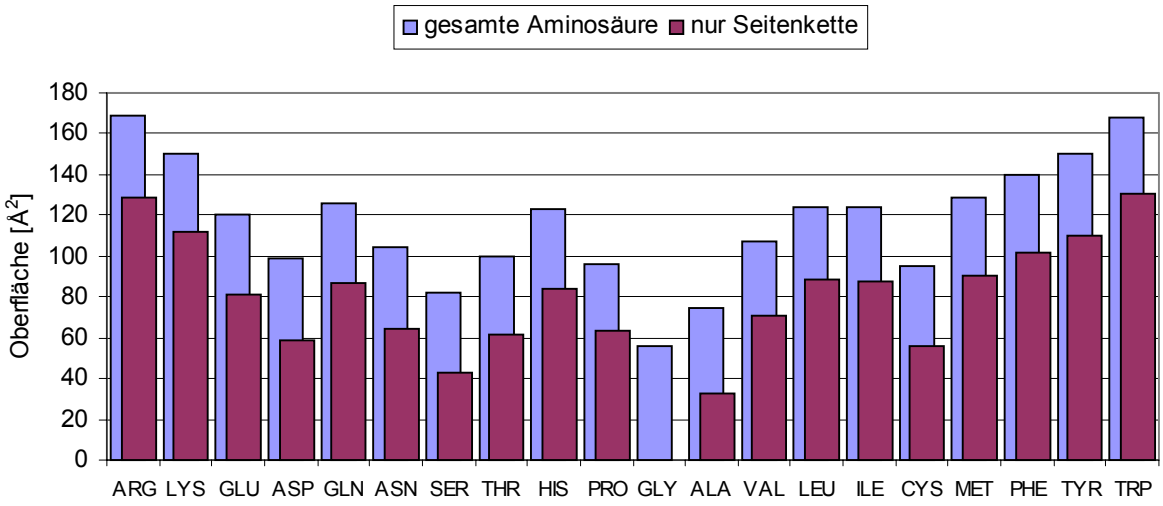


Diagramm 5.17: Oberflächen A_i^0 der 20 proteinogenen Aminosäuren im Tripeptid Glycin-X-Glycin. Für jede Aminosäure ist sowohl der Flächeninhalt der gesamten Aminosäure als auch nur der Flächeninhalt der Seitenkette an der Molekülоberfläche dargestellt.

In Diagramm 5.17 sind die Oberflächen der Aminosäuren in den Tripeptiden dargestellt. Deutlich sind die verschiedenen Größen der Aminosäuren zu erkennen. Glycin, welches keine Seitenkette besitzt, hat die kleinste Oberfläche an der Molekülaußenseite (56 Å^2). Arginin und Tryptophan hingegen besitzen die größte Oberfläche (169 bzw. 168 Å^2). Die Differenz der Gesamtoberfläche der Aminosäure und der Oberfläche der Seitenkette ist bei allen Aminosäuren ähnlich (ca. 40 Å^2). Das bedeutet, man kann die Oberfläche der Aminosäuren in zwei Teilbereiche zerlegt betrachten: Der erste Teil ist die Oberfläche, die durch die Atome des Proteinrückgrates gebildet wird. Diese Fläche ist für alle Aminosäuren etwa gleich groß, und dieser Molekülteil hat auch ähnliche molekulare Eigenschaften. Der zweite Oberflächenteil wird durch die Seitenkette der Aminosäure bestimmt, diese hat sehr verschiedene Größen und molekulare Eigenschaften. Einzige Ausnahme ist hierbei Prolin, das eine Ringstruktur besitzt (der Unterschied zwischen Gesamtfläche und Seitenkette beträgt ca. 30 Å^2).

5.5.2 Oberfläche von einzelnen Aminosäuren in Proteinen

Diagramm 5.18 zeigt die gemittelten Flächeninhalte der Aminosäuren bzw. Aminosäuren-seitenketten an der Proteinoberfläche aller untersuchten Proteine, und Tabelle 9.6 enthält die entsprechenden Zahlenwerte und Standardabweichungen.

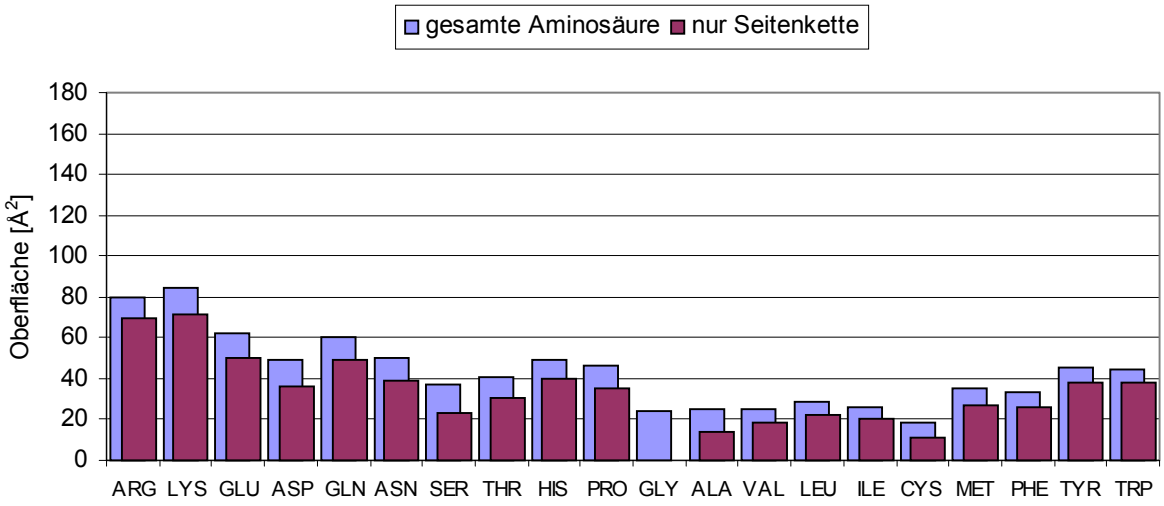


Diagramm 5.18: Gemittelte Oberflächen A_i der 20 proteinogenen Aminosäuren bzw. Aminosäurenseitenketten an der molekularen Proteinoberfläche.

Lysin und Arginin besitzen die größte mittlere Oberfläche (84 bzw. 80 $Å^2$) an den Proteinaußenseiten. Die mittlere Oberfläche der Aminosäuren an den Proteinoberflächen beträgt etwa die Hälfte der Flächen in den Tripeptiden. Die Differenz der Gesamtaminosäurefläche und Seitenkettenfläche beträgt nur noch 10 $Å^2$, d.h. etwa ein Viertel im Vergleich zum Tripeptid. Die Aminosäurenseitenkette nimmt an der Proteinoberfläche relativ zur gesamten Aminosäure mehr Platz als im Tripeptid bzw. einer gestreckten Polypeptidkette ein.

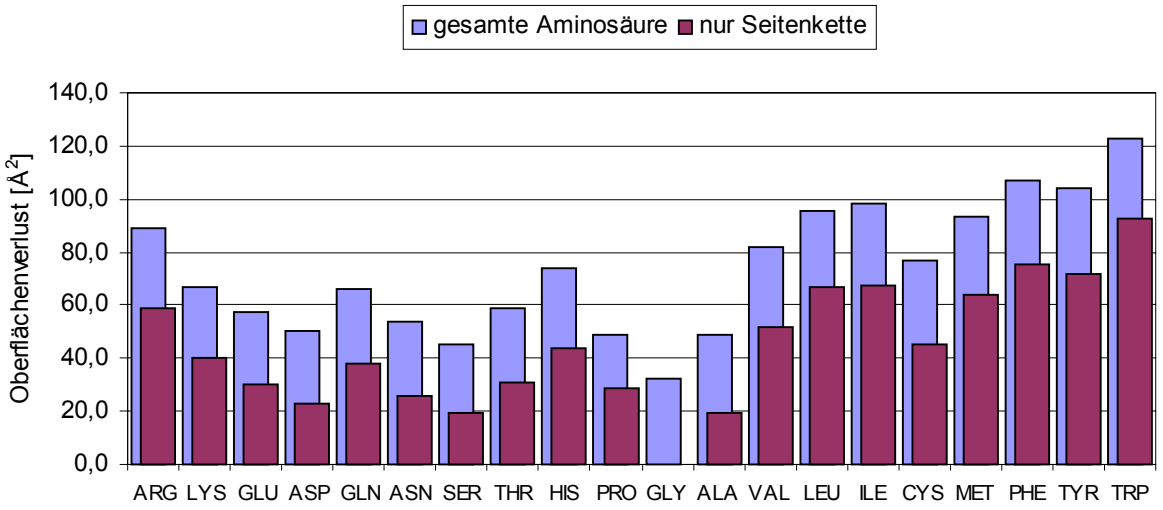


Diagramm 5.19: Oberflächenverluste $A_i^0 - A_i$ der 20 proteinogenen Aminosäuren. Sowohl der Oberflächenverlust der gesamten Aminosäure als auch nur von der Seitenkette ist dargestellt.

Die Differenz der Flächen $A_i^0 - A_i$, d.h. der Unterschied der Aminosäurenoberflächen in den Tripeptiden und Proteinen, gibt den Oberflächenverlust der Aminosäuren durch die Proteinfaltung wieder (siehe Diagramm 5.19). Die Aminosäuren verlieren zwischen 30 bis 120 $Å^2$ Oberfläche durch die Faltung der Polypeptidkette, wobei die Abnahme bei hydrophoben Aminosäuren (Alanin, Valin, Leucin, Isoleucin, Methionin, Phenylalanin, Tyrosin und Cystein) größer ist als bei hydrophilen Aminosäuren (mit Ausnahme von Arginin). Tryptophan verliert mit knapp 120 $Å^2$ am meisten Oberfläche. Dies ist im Einklang mit dem Anteil der Aminosäuren an den Proteinoberflächen (siehe Kapitel 5.4.2). Neben den Absolutwerten sind auch die relativen Oberflächenverluste $f_{rel,i}$ (Gleichung 4.3 in Kapitel 4.2.5) der einzelnen Aminosäuren interessant. In Diagramm 5.20 sind die relativen Oberflächenverluste der 20 proteinogenen Aminosäuren zusammengestellt.

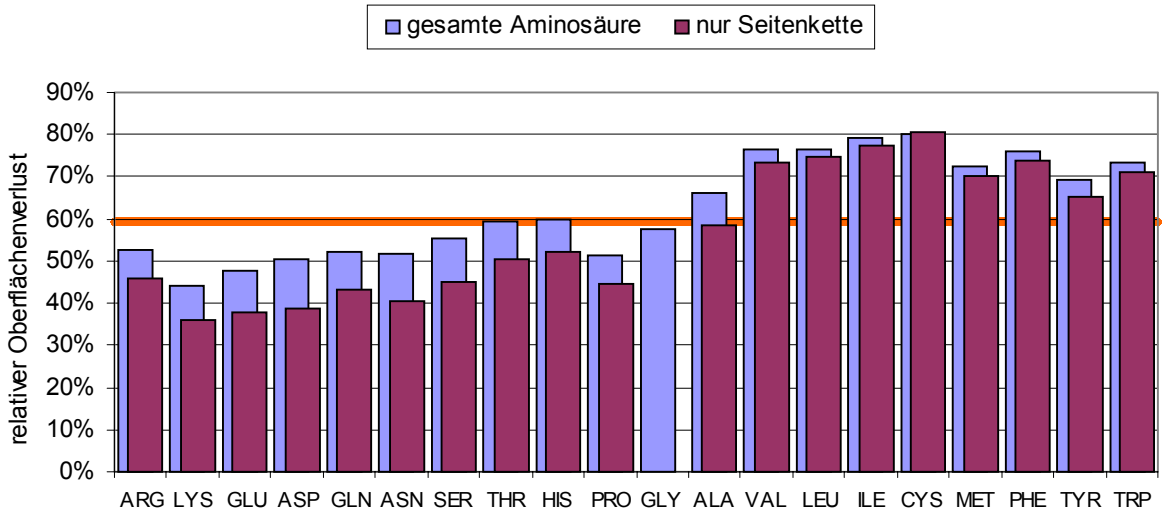


Diagramm 5.20: Relative Oberflächenverluste $f_{rel,i}$ von Aminosäuren und Aminosäurenseitenketten.

Der Unterschied zwischen hydrophilen und hydrophoben Aminosäuren ist deutlich sichtbar. Der relative Oberflächenverlust der hydrophoben Aminosäuren an der Proteinoberfläche beträgt 70% bis 80%. D.h. sie besitzen im Vergleich zum gestreckten Polypeptid (siehe Werte der Tripeptide) nur noch zwischen 20% bis 30% ihrer Oberfläche. Cystein verliert relativ gesehen am meisten Oberfläche (80%). Der relative Oberflächenverlust der hydrophilen Aminosäuren ist unter 60%. Die Aminosäuren Lysin, Glutaminsäure, Asparaginsäure (geladene Aminosäuren) und Prolin haben den niedrigsten relativen Oberflächenverlust (ca. 50%). Durch die 60%-Grenze lassen sich die Aminosäuren in hydrophile und hydrophobe Aminosäuren unterteilen. Der relative Oberflächenverlust von

Alanin liegt an dieser Grenze, die gesamte Aminosäure verliert mehr als 60%, die Seitenkette weniger als 60% durch die Faltung der Polypeptidkette. Alanin hat eine Methylgruppe als Seitenkette und wird meistens zu den hydrophoben Aminosäuren gezählt, da keine polaren Atome in seiner Seitenkette vorhanden sind. Die Methylgruppe ist jedoch im Gegensatz zu den Seitenketten der anderen hydrophoben Aminosäuren klein. Alanin besitzt nach Glycin die kleinste Seitenkette (in Bezug auf Oberfläche, Volumen und Anzahl der Atome) aller 20 proteinogenen Aminosäuren. Im Gegensatz zu den anderen Aminosäuren (mit Ausnahme von Glycin) nimmt die Seitenkette von Alanin weniger Platz an der Oberfläche des gestreckten Polypeptids ein als die Atome des Rückgrates (siehe Diagramm 5.17). Alanin ist somit eher mit Glycin als mit den anderen hydrophoben Aminosäuren vergleichbar.

Diagramm 5.21 zeigt die Verteilung der mittleren Oberfläche A_i der 20 proteinogenen Aminosäuren an der Gesamtoberfläche der Proteine. Die Darstellung ist aufgeteilt in fünf Einzeldiagramme (a-e) mit den Verteilungskurven von Aminosäuren mit ähnlichen Eigenschaften bzw. Seitenketten. In den Diagrammen ist erneut der bereits beschriebene Unterschied zwischen hydrophilen und hydrophoben Aminosäuren sichtbar. Die Verteilungen der Flächenanteile von hydrophilen Aminosäuren haben ein Maximum bei einem mittleren Flächeninhalt (50-100 $Å^2$), und nur wenige dieser Aminosäuren haben eine Fläche unter 5 $Å^2$ an der Proteinoberfläche. Die hydrophoben Aminosäuren zeigen den entgegengesetzten Trend: Das Maximum liegt bei sehr geringen Flächeninhalten (0-20 $Å^2$). Hydrophile Aminosäuren ragen also bevorzugt aus den Proteinen heraus und möchten möglichst viel mit dem Solvens in Kontakt treten. Hydrophobe Aminosäuren ziehen sich hingegen in das Proteininnere zurück. Diese Ergebnisse stimmen mit den Daten in Kapitel 5.4.2 überein. Alanin nimmt unter den hydrophoben Aminosäuren, wie oben bereits erwähnt, eine Sonderstellung ein. Die Verteilungskurve von Alanin ähnelt mehr der Flächenverteilung von Glycin als den anderen hydrophoben Aminosäuren wie z.B. Valin oder Leucin. Leichte Abweichungen in den Verteilungskurven sind auch beim Tryptophan und Tyrosin erkennbar. Beide gelten als hydrophobe Aminosäuren, jedoch ist das Maximum der Flächenverteilung im Gegensatz zu den anderen hydrophoben Aminosäuren nicht bei minimalem Flächeninhalt, sondern bei 15 $Å^2$. Tryptophan und Tyrosin enthalten in ihren Seitenketten polare Sauerstoff- bzw. Stickstoffatome. Dies ist für den Unterschied in den Verteilungskurven verantwortlich. In allen Verteilungskurven entspricht der Grenzwert, bei dem die Verteilungskurve auf 0% Häufigkeit abfällt, der im Tripeptid bestimmten Oberfläche der jeweiligen Aminosäure.

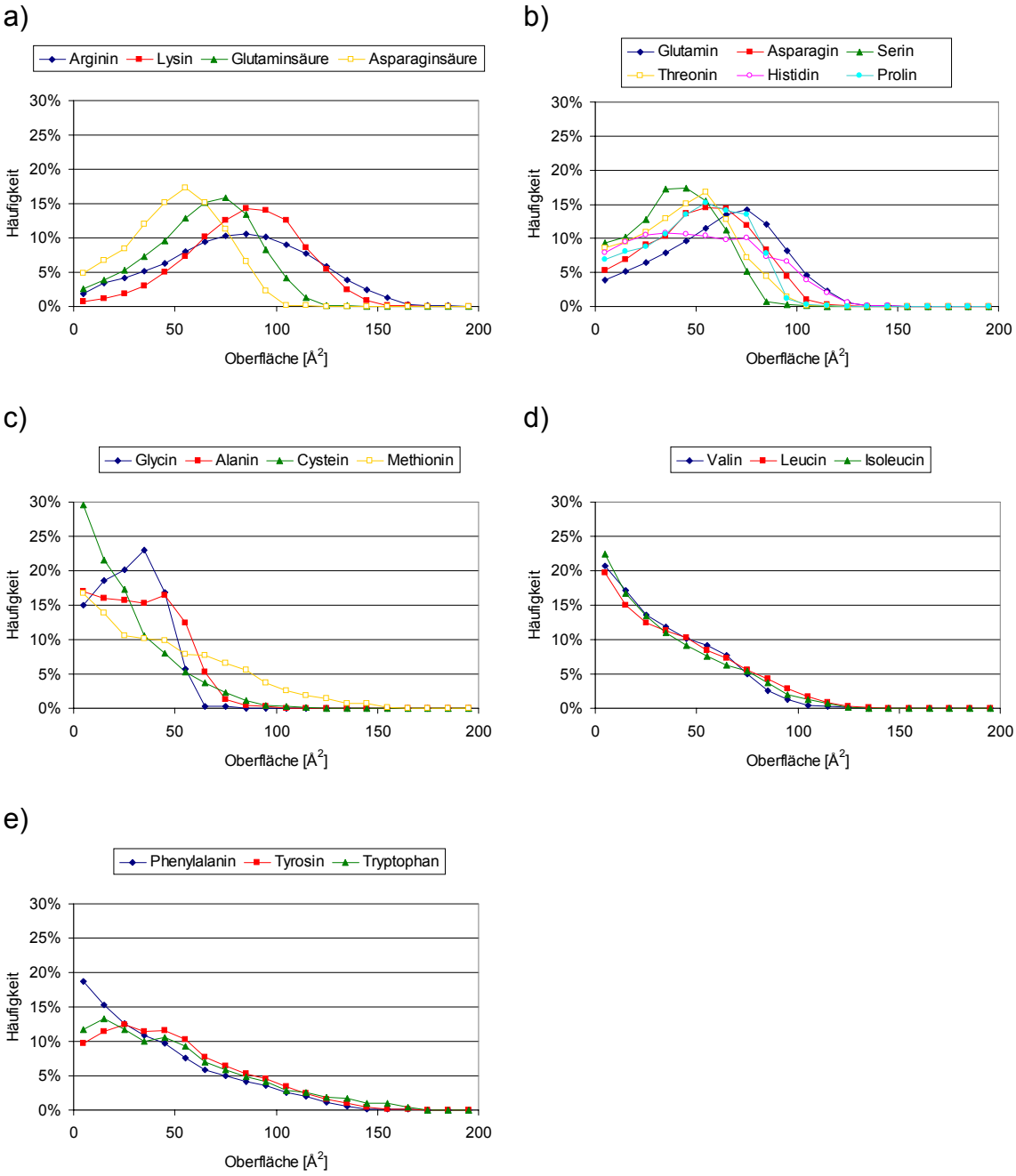


Diagramm 5.21: Größenverteilung (Gesamtfläche der Aminosäuren) der 20 proteinogenen Aminosäuren an der Gesamtoberfläche von Proteinen.

5.5.3 Oberfläche von einzelnen Aminosäuren in Bindungsbereichen

Besonderes Interesse gilt in dieser Arbeit den Bindungsbereichen von Proteinkomplexen, deshalb werden im Folgenden die mittleren Flächenanteile (Flächeninhalte A_i) der verschiedenen Aminosäuretypen auf der Proteinoberfläche in Bindungsbereichen bestimmt und mit den an der gesamten Oberfläche ermittelten Werten verglichen.

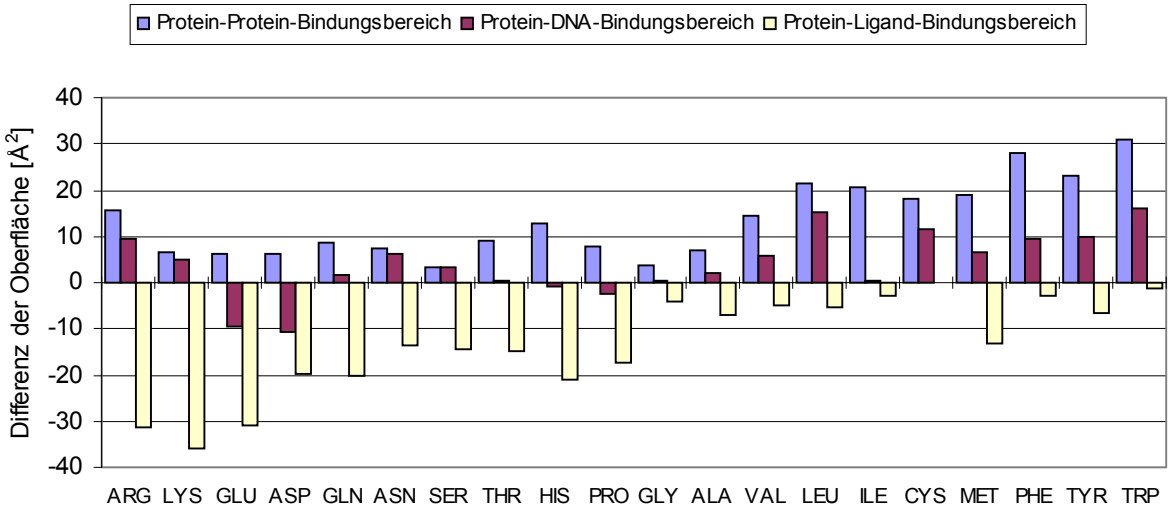


Diagramm 5.22: Analyse der mittleren Oberflächen A_i von Aminosäuren in verschiedenen Bereichen der molekularen Oberfläche. Für jeden Bindungsbereich (Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereich) ist die Differenz $A_{i,Bindungsbereich} - A_{i,Gesamtoberfläche}$ der mittleren Oberfläche einzelner Aminosäuren im Bindungsbereich und an der gesamten Proteinoberfläche aufgetragen.

In Diagramm 5.22 sind die Differenzen der mittleren Aminosäureoberflächen in Bindungsbereichen (Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereiche) und an der Gesamtoberfläche aufgetragen. Es fällt auf, daß die Aminosäuren im Protein-Protein- und Protein-DNA-Bindungsbereich der Proteinoberfläche mehr Platz einnehmen als im nichtbindenden Bereich der molekularen Oberfläche. Besonders ausgeprägt ist dieses Verhalten in den Protein-Protein-Bindungsbereichen. Diese größere Oberfläche der Aminosäuren bedeutet, daß diese in den Protein-Protein-Bindungsbereichen weiter aus den Proteinen herausragen. Dadurch können sie etwaige Leerräume zwischen den Proteinen in Proteinkomplexen ausfüllen und so zu einer engeren bzw. festeren Bindung der Molekülpartner führen. Weiterhin wird dadurch die Kontaktfläche pro Aminosäure im Bindungsbereich maximiert. Die aromatischen Aminosäuren Phenylalanin, Tyrosin und Tryptophan haben im Vergleich zur gesamten Proteinoberfläche die größte Steigerung der Oberfläche pro Aminosäure in Protein-Protein-Bindungsbereichen. Die aromatischen Systeme können durch den höheren Anteil an der molekularen Oberfläche stärkere Wechselwirkungen im Bindungsbereich ausüben. In den Protein-DNA-Bindungsbereichen unterscheiden sich die Oberflächenanteile der negativ geladenen Aminosäuren Asparagin- und Glutaminsäure von den anderen Aminosäuren. Ihre Oberflächenanteile sind im DNA-Bindungsbereich deutlich niedriger als an der nichtbindenden Proteinoberfläche. Die Oberflächenanteile

aller anderen Aminosäuren sind entweder fast unverändert oder erhöht. Dies stimmt mit den bereits im vorhergehenden Kapitel 5.4.3.3 beschriebenen Ergebnissen überein. Im DNA-Bindungsbereich sind negativ geladene Aminosäuren weniger vorhanden oder nehmen weniger Fläche an der Bindungsoberfläche ein, so werden ungünstige abstoßende Wechselwirkungen zu den negativ polarisierten DNA-Molekülen vermieden.

Im Protein-Ligand-Bindungsbereich ist ein anderer Trend zu beobachten. Hier sind die mittleren Flächen aller Aminosäurentypen an der Bindungsoberfläche kleiner als an der nichtbindenden Proteinoberfläche. Die hydrophilen Aminosäuren haben wiederum mehr Oberfläche als die hydrophoben Aminosäuren. Diese allgemein kleineren Flächen der Aminosäuren in Protein-Ligand-Bindungsstellen lassen sich sehr einfach erklären, wenn man die Form der Ligandbindungsbereiche analysiert. Ligandbindungsstellen befinden sich bevorzugt in Spalten, Taschen oder Höhlen der molekularen Oberfläche (siehe auch Kapitel 5.7.4). Durch diese konkave Struktur der Oberfläche resultiert automatisch eine Verringerung des Oberflächenbereiches jeder Aminosäure (siehe Abbildung 5.1).

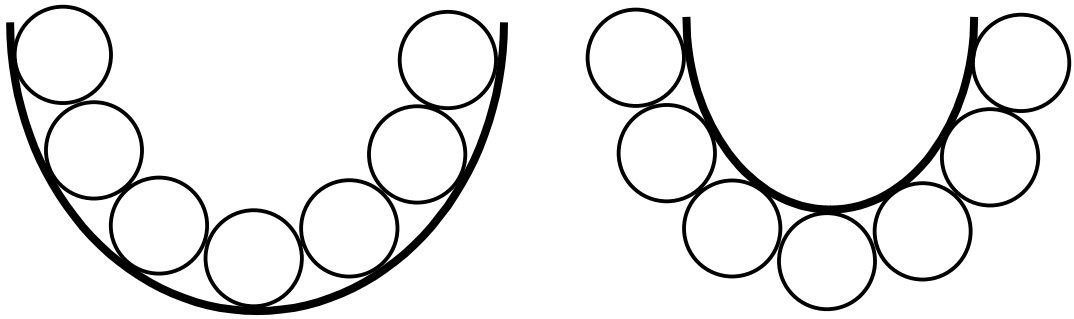


Abbildung 5.1: Atome bzw. Aminosäuren an der Oberfläche von konvexen (links) bzw. konkaven (rechts) Bereichen der molekularen Oberfläche.

5.6 Kontaktwechselwirkungen in Protein-Protein-Komplexen

In den vorhergehenden Kapiteln wurde die relative Häufigkeit der 20 proteinogenen Aminosäuren an der Proteinoberfläche und in deren Bindungsbereichen analysiert. Nun stellt sich die Frage, ob diese Verteilung Einfluß auf die Kontakte von Aminosäuren in Protein-Protein-Bindungsbereichen hat. Um dies zu beantworten, werden die Kontaktmatrizen $\mathbf{M}_{\text{Kontakt,beob.}}$ (siehe Kapitel 4.2.6) für die vier Datensätze berechnet und untersucht. Zur Illustration der Matrizen werden diese als Gitter mit 20x20 Feldern dargestellt (siehe Diagramm 5.23). Die Farbe oder der Grauwert des Feldes gibt die relative Kontaktfläche (in Bezug zur Gesamtkontaktfläche des Protein-Protein-Bindungsbereiches) wieder. Die Matrizen sind spiegelsymmetrisch entlang der Hauptdiagonalen.
Beim Vergleich der Kontaktmatrizen in den vier Datensätzen werden teilweise starke Unterschiede deutlich. Es gibt kein allgemein gültiges Kontaktmuster der Aminosäuren in den Bindungsbereichen der Protein-Protein-Komplexe. Es sind jedoch einige Gemeinsamkeiten zu erkennen:

1. Es werden bevorzugt Salzbrücken zwischen den positiv geladenen Aminosäuren Arginin(R) und Lysin(K) und den negativ geladenen Aminosäuren Asparaginsäure(E) und Glutaminsäure(D) gebildet.
2. Leucin(L)-Leucin(L)-Kontakte sind häufig zu finden.
3. In den Diagrammen sind die Disulfidbrücken (Cystein(C)-Cystein(C)-Kontakte) gut sichtbar.

Die Protein-Protein-Bindungsbereiche im Antigen-Antikörper-Datensatz unterscheiden sich stark von den Bindungsbereichen in den anderen Datensätzen. Tyrosin-Kontakte spielen hier eine große Rolle, was auf die hohe relative Häufigkeit von Aminosäuren mit aromatischen Seitenketten (Tyrosin, Phenylalanin und Tryptophan) in den Protein-Protein-Bindungsbereichen der Antigen-Antikörper-Komplexe zurückgeführt werden kann (siehe Diagramm 5.14 und Diagramm 9.1 im Anhang).

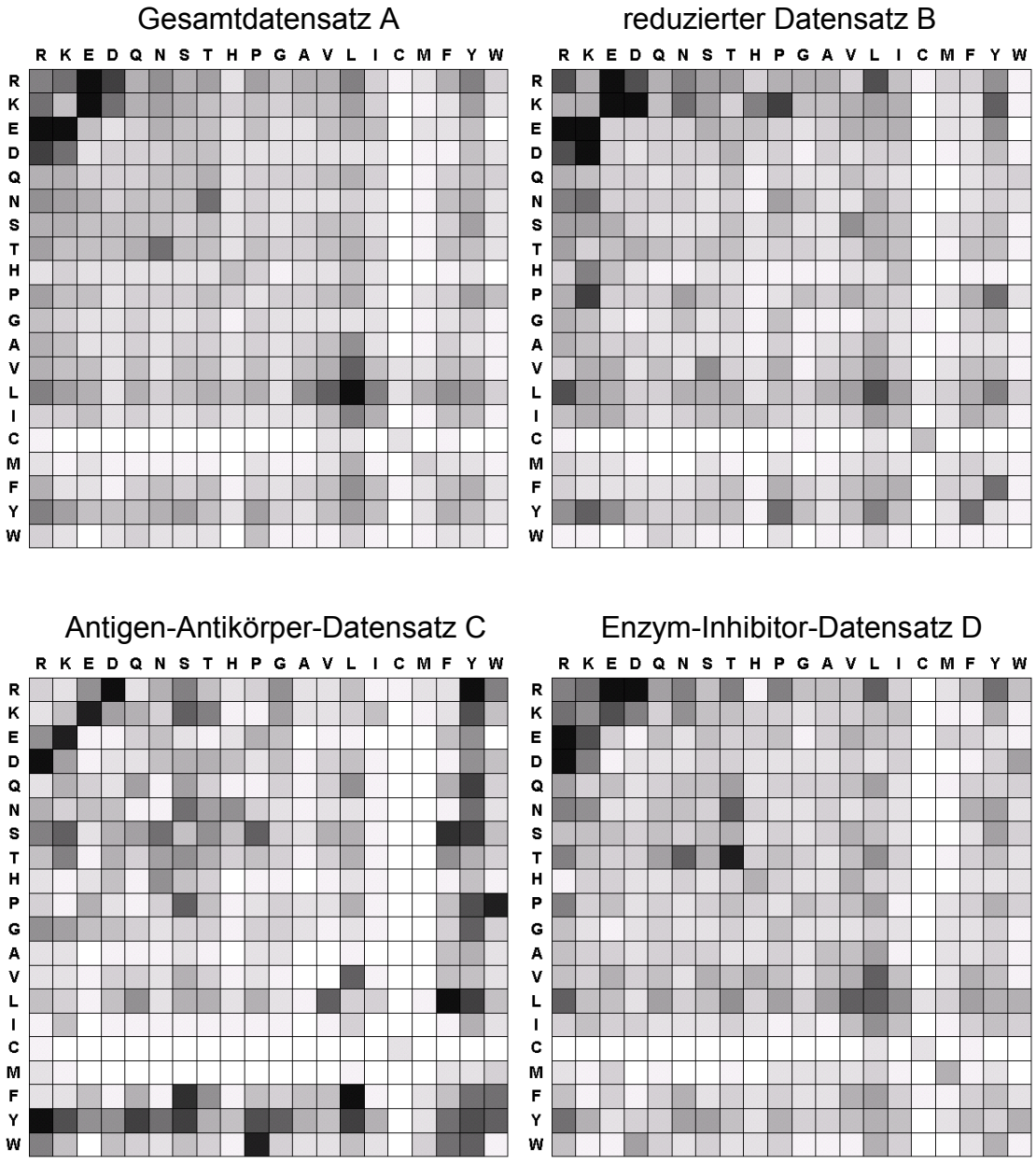


Diagramm 5.23: Relative Kontaktfläche von Aminosäuren zwischen Proteinen in den Datensätzen A-D (weiß: 0,0%; schwarz: 1,0%). Die Reihen und Spalten sind entsprechend der Ein-Buchstaben-Schreibweise der proteinogenen Aminosäuren beschriftet (Tabelle 9.1).

Die Anzahl von Kontakten zwischen bestimmten Aminosäurespezies ist von der Häufigkeit der jeweiligen Aminosäure im Bindungsbereich der Proteinkomplexe abhängig. Um diese Abhängigkeit bei der Untersuchung zu berücksichtigen, werden die theoretischen Kontaktmatrizen $\mathbf{M}_{\text{Kontakt,theoretisch}}$ mit Hilfe der relativen Häufigkeiten der Aminosäuren berechnet und von den beobachteten Matrizen abgezogen (siehe Kapitel 4.2.6). Die Differenzmatrizen sind in Diagramm 5.24 und die theoretischen Kontaktmatrizen im Diagramm 9.1 im Anhang dargestellt.

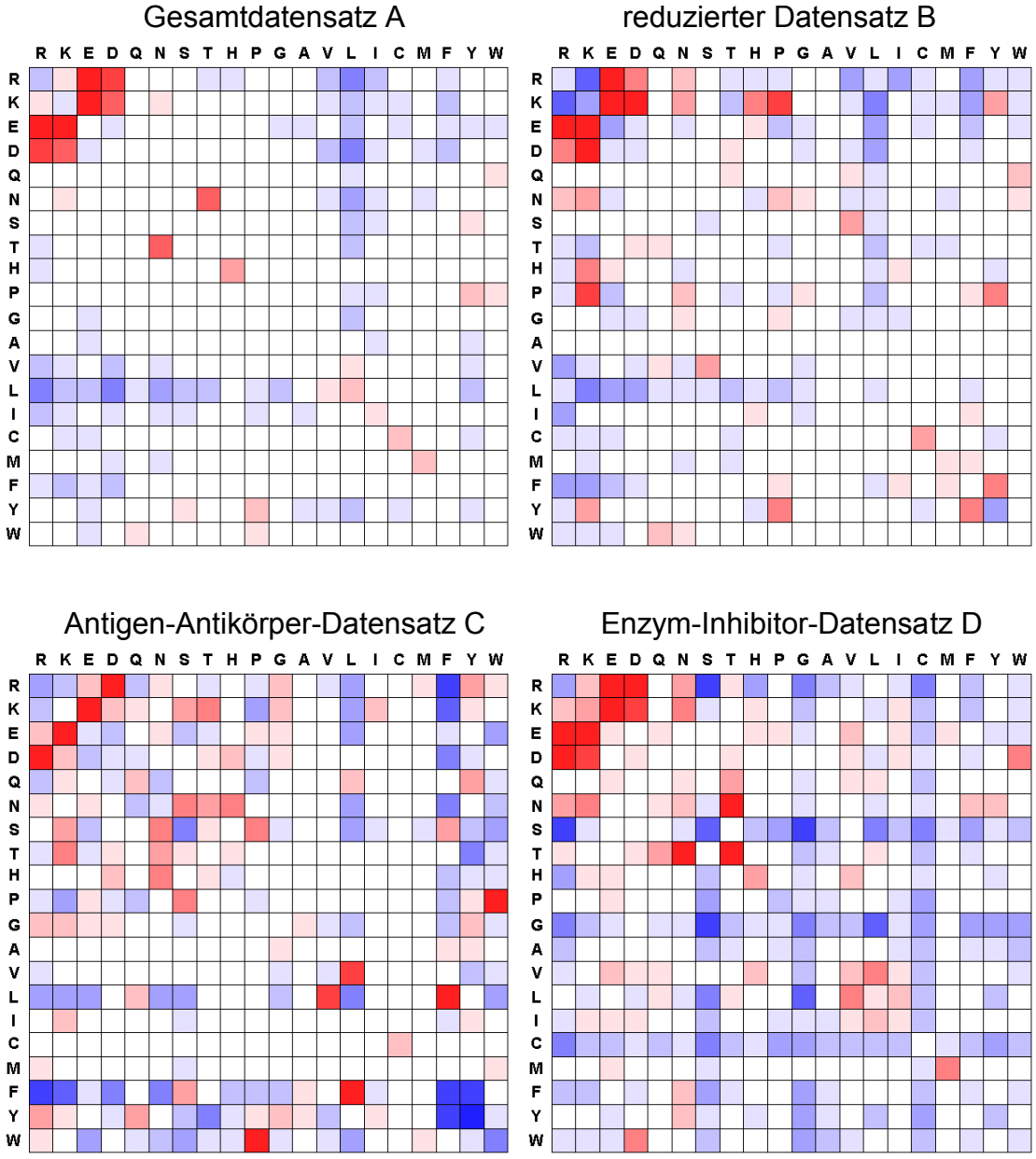


Diagramm 5.24: Differenz der beobachteten und theoretischen relativen Kontaktflächen der Aminosäuren zwischen Proteinen in den Datensätzen A-D (blau: -0,5%; weiß 0,0%; rot: +0,5%).

Die Differenzmatrizen zeigen, welche Aminosäurekontakte abweichend von den relativen Häufigkeiten der Aminosäuren an der molekularen Oberfläche im Bindungsbereich bevorzugt oder benachteiligt ausgebildet werden. Die vier Diagramme unterscheiden sich deutlich. Je weniger Protein-Protein-Bindungsbereiche im Datensatz vorhanden sind (Tabelle 5.5), desto mehr weicht die gemittelte, beobachtete Kontaktmatrix $\mathbf{M}_{\text{Kontakt,beob.}}$ von der berechneten Matrix $\mathbf{M}_{\text{Kontakt,theoretisch}}$ ab. Dies ist wiederum ein Indiz dafür, daß die Kontakte der Aminosäuren in den Bindungsbereichen der Protein-Protein-Komplexe sehr verschieden sind und spezifisch von der Art und Funktion der Protein-Protein-Wechselwirkung abhängen. Jedoch lassen sich aus der Differenzmatrix der Komplexe des Gesamtdatensatzes einige Tendenzen erkennen, die teilweise auch in den anderen Datensätzen sichtbar sind. Aminosäurekontakte zwischen positiv und negativ geladenen Aminosäuren (Salzbrücken) sind deutlich häufiger, als die relative Häufigkeit dieser Aminosäuren an der molekularen Oberfläche vermuten läßt. Dies gilt auch für Kontakte zwischen Serin und Threonin. Diese beiden Aminosäuren sind zwar im Vergleich zu den anderen Aminosäuren recht klein, besitzen aber Hydroxylgruppen und können damit Wasserstoffbrückenbindungen zwischen den Proteinen ausbilden. Die hydrophoben Kontakte zwischen Leucin und Leucin sind auch häufiger vorhanden, als aufgrund der Häufigkeit vorhergesagt werden würde. Das gilt ebenso für die Cystein-Cystein-Kontakte, die Disulfidbrücken zwischen den Proteinen ausbilden. Weniger Kontakte als berechnet werden zwischen den hydrophilen Aminosäuren und Leucin gebildet. Zwischen den Seitenketten dieser Aminosäuren sind keine hydrophoben oder polaren Wechselwirkungen möglich, die zur Stabilisierung der Komplexe beitragen können. Eingeschränkt gilt das auch für die Kontakte von Valin und Isoleucin mit den hydrophilen Aminosäuren.

5.7 Untersuchung der molekularen Proteinoberflächen

Zur Analyse der lokalen molekularen Eigenschaften der Proteine werden diese berechnet und auf die Proteinoberflächen projiziert (siehe Kapitel 3.2), die molekularen Oberflächen in Teilstücke eingeteilt und anschließend die Eigenschaftswerte über all diese Teiloberflächen gemittelt (Kapitel 4.1.5). Insgesamt ergeben sich so im Gesamtdatensatz A mehr als eine Million Teiloberflächen, welche die lokalen Eigenschaften der Proteine repräsentieren. Tabelle 5.13 enthält die Anzahl der Teiloberflächen in den vier Datensätzen. Im Mittel wird eine Proteinoberfläche in 97 überlappende Teilstücke unterteilt, und die durchschnittliche Größe eines dieser Oberflächenteilstücke beträgt 261,7 $Å^2$. Das entspricht etwa der Fläche, die 4-6 Aminosäuren an der Proteinaußenseite einnehmen (siehe Tabelle 9.6 im Anhang). Je nach Datensatz befinden sich etwa 11-17% der Oberflächensegmente in Protein-Protein-Bindungsbereichen der Proteinoberfläche. Unter 1% der Segmente binden DNA bzw. Ligandmoleküle. Über 80% der Teiloberflächen bilden keine Bindung zu anderen Molekülen aus.

Tabelle 5.13: Anzahl der Teiloberflächen in den vier Datensätzen unterteilt nach Gesamtoberfläche, Protein-Protein-, Protein-DNA- und Protein-Ligand-Bindungsbereich.

Datensatz	Gesamtoberfläche	Protein-Protein-Bindungsbereich	Protein-DNA-Bindungsbereich	Protein-Ligand-Bindungsbereich
A	1255853	147784 ≙11,8%	3273 ≙0,3%	10192 ≙0,8%
B	75429	7741 ≙10,3%	244 ≙0,3%	588 ≙0,8%
C	56646	9442 ≙16,7%	3 ≙0,0%	218 ≙0,4%
D	14670	2459 ≙16,8%	0 ≙0,0%	102 ≙0,7%

Die Mittelwerte der auf die Teiloberflächen projizierten Eigenschaften in den vier verschiedenen Datensätzen sind in den Tabellen 9.8 bis 9.11 im Anhang zusammengestellt. In den folgenden Kapiteln werden die einzelnen molekularen Eigenschaften anhand der Mittelwerte und Verteilungskurven analysiert. Die Daten werden dabei getrennt nach den vier verschiedenen Oberflächenbereichen (gesamte Proteinoberfläche, Protein-Protein-Bindungsbereich, Protein-DNA-Bindungsbereich oder Protein-Ligand-Bindungsbereich) und den vier Datensätzen A-D betrachtet. In den Protein-DNA-Bindungsbereichen beschränken sich die Untersuchungen der Mittelwerte wie in den vorhergehenden Kapiteln auf die Datensätze A und B.

5.7.1 Elektrostatisches Potential

Das auf die Teiloberflächen projizierte und gemittelte elektrostatische Potential unterscheidet sich in den verschiedenen Bereichen der Proteinoberfläche (Diagramm 5.25). Der Mittelwert des elektrostatischen Potentials von Teiloberflächen der gesamten Proteinoberfläche und in den Protein-Protein-Bindungsbereichen ist in allen vier Datensätzen erwartungsgemäß fast null. Die Partialladungen der Aminosäuren an der Proteinaußenseite und in den Protein-Protein-Bindungsbereichen gleichen sich in der Summe aus. Weder positiv noch negativ geladene Aminosäuren sind in diesen Bereichen bevorzugt. In Protein-DNA-Bindungsbereichen sind hingegen positiv geladene bzw. polarisierte Aminosäuren häufiger vorhanden. Das durchschnittliche elektrostatische Potential an diesen Oberflächensegmenten beträgt ca. 50 kcal/(mol·e). Dies stimmt mit den Ergebnissen aus Kapitel 5.4.3.3 überein. Dort wurde gezeigt, daß die positiv geladenen Aminosäuren Arginin und Lysin im DNA-Bindungsbereich der Proteine sehr oft vorkommen, während die elektronegativ geladenen Aminosäuren Glutaminsäure und Asparaginsäure nur selten vorhanden sind. Die Protein-Ligand-Bindungsbereiche sind elektropositiv polarisiert, jedoch sehr viel schwächer als die DNA-Bindungsstellen (nur ca. 10 kcal/(mol·e)). Die Komplexe des Datensatzes C (Antikörper-Antigen-Komplexe) bilden hier eine Ausnahme: Der Mittelwert beträgt fast null und ähnelt mehr dem auf die Gesamtoberfläche oder die Protein-Protein-Bindungsregionen projizierten elektrostatischen Potential.

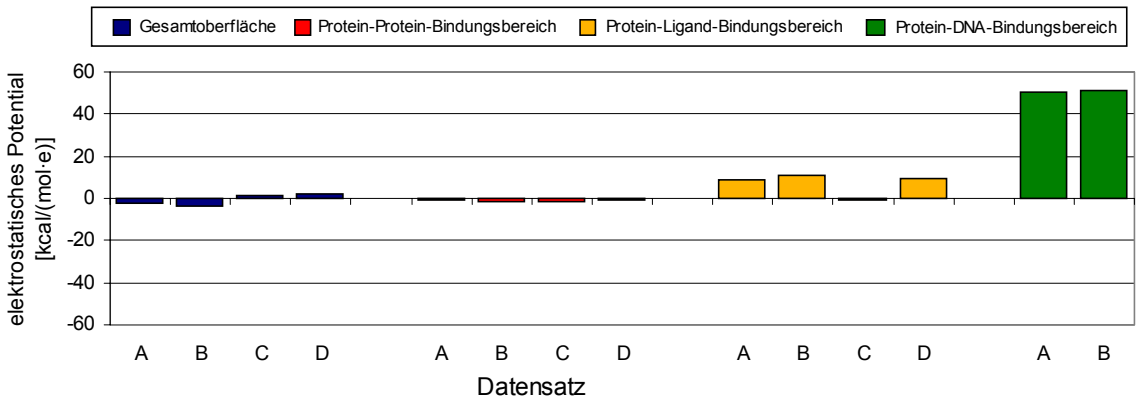


Diagramm 5.25: Elektrostatisches Potential gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente. Die Mittelwerte sind nach den Oberflächenbereichen (Bindungsregionen) gruppiert und eingefärbt. Innerhalb der vier Gruppen sind die Werte der verschiedenen Datensätzen A-D jeweils gegenübergestellt.

Diagramm 5.26 zeigt die Verteilung der Mittelwerte des elektrostatischen Potentials in den vier unterschiedlichen Oberflächenbereichen der Proteine des Gesamtdatensatzes A. Wenn die gesamte Proteinoberfläche betrachtet wird, ähnelt die Verteilung des elektrostatischen Potentials einer Glockenkurve mit dem Maximum am Nullpunkt. Die Verteilung der Potentialwerte in den Protein-Protein-Bindungsbereichen ist sehr ähnlich, jedoch ist das Maximum am Nullpunkt etwas höher, d.h. in den Protein-Protein-Bindungsbereichen der Proteinkomplexe befinden sich weniger Regionen mit extrem hohen bzw. extrem niedrigen Potentialwerten und mehr Teiloberflächen mit geringen Absolutwerten des elektrostatischen Potentials. Der Verlauf der Verteilungskurve für die DNA-Bindungsstellen zeigt deutlich den schon erwähnten Trend zu elektropositiv geladenen Aminosäuren. Die Kurve ähnelt der Verteilung über die Gesamtoberfläche, ist aber um 50 kcal/(mol·e) in den elektropositiven Bereich verschoben. Eine leichte Verschiebung ist auch bei den Protein-Ligand-Bindungsbereichen sichtbar. Im Vergleich zur Gesamtoberfläche ist der Anteil an Oberflächensegmenten mit einem Potential zwischen 40 und 100 kcal/(mol·e) erhöht und der Anteil der Segmente mit Werten zwischen 0 und –40 kcal/(mol·e) erniedrigt.

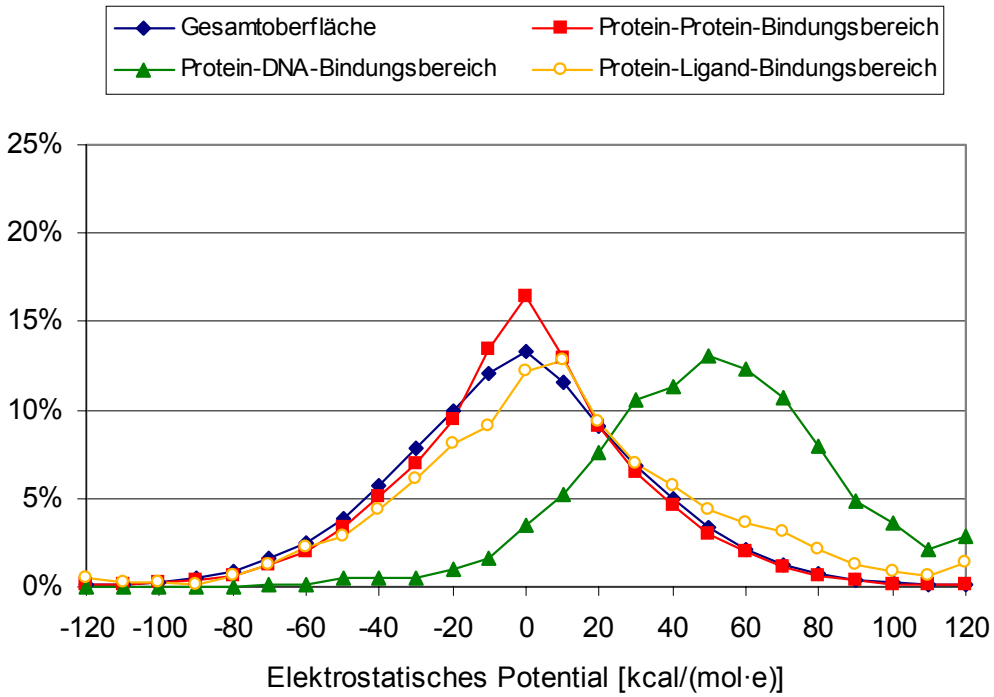


Diagramm 5.26: Gemitteltes elektrostatisches Potential von Oberflächensegmenten der Proteine des Gesamtdatensatzes A: Die Kurven verdeutlichen die Verteilung des elektrostatischen Potentials an den vier verschiedenen Oberflächenbereichen der Proteine (gesamte Oberfläche, Protein-Protein-, Protein-DNA- oder Protein-Ligand-Bindungsregionen).

5.7.2 Lokale Lipophilie

Die auf die Teiloberflächen projizierten Werte der lokalen Lipophilie (Diagramm 5.27) sind im Mittel negativ, d.h. im hydrophilem Bereich der hier verwendeten Lipophilie-Skala (siehe Kapitel 3.2.2). Wie in den vorhergehenden Kapiteln erläutert wurde, besteht die Außenseite der untersuchten Proteinstrukturen zum großen Teil aus hydrophilen Aminosäuren, die attraktive Wechselwirkungen zum wäßrigen Lösungsmittel ausbilden können. In der Proteindatenbank sind nur sehr wenige Strukturen von Proteinen mit großen hydrophoben Außenbereichen wie z.B. Transmembranproteine erfaßt. Die gemittelten Werte der Lipophilie sind an der Gesamtoberfläche (alle Bereiche der Oberfläche) am niedrigsten (d.h. am meisten hydrophil). Dabei gibt es keine signifikanten Unterschiede zwischen den vier Datensätzen A-D. Die Oberflächensegmente in den Protein-DNA-Bindungsbereichen haben ähnlich niedrige Mittelwerte und sind entsprechend hydrophil. Die Protein-Protein- und Protein-Ligand-Bindungsbereiche sind hingegen deutlich hydrophober. Die Ligandbindungsstellen sind dabei etwas hydrophober als die Protein-Protein-Bindungsstellen. In diesen Oberflächenbereichen gibt es auch Unterschiede zwischen den Datensätzen. Die Mittelwerte der Datensätze A und B sind sehr ähnlich. Datensatz C (Antigen-Antikörper-Komplexe) weicht jedoch sowohl im Protein-Protein- als auch im Protein-Ligand-Bindungsbereich ab. In beiden Fällen werden in den Antigen-Antikörper-Komplexen hydrophobere Mittelwerte beobachtet. Im Datensatz D (Enzym-Inhibitor-Komplexe) ist ein entgegengesetzter Trend zu beobachten. Die Protein-Ligand- und Protein-Protein-Bindungsbereiche sind hydrophiler als im Gesamtdatensatz. Die Lipophiliewerte beider Bereiche sind im Gegensatz zu den anderen Datensätzen fast identisch.

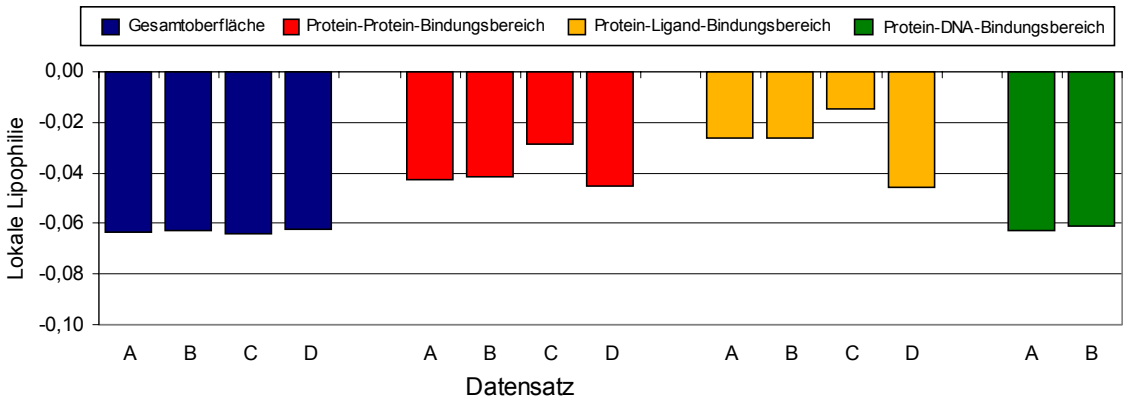


Diagramm 5.27: Lokale Lipophilie gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente.

Die Verteilungskurven der lokalen Lipophiliewerte (Diagramm 5.28) verdeutlichen noch einmal die schon erwähnten Beobachtungen: Die Proteingesamtoberfläche sowie die Protein-DNA-Bindungsbereiche sind am stärksten hydrophil. Die Verteilungskurve der DNA-Bindungsbereiche verläuft etwas höher und steiler als die der Gesamtoberfläche. An der Gesamtoberfläche sind etwas mehr Oberflächensegmente mit extrem hoher bzw. niedriger Lipophilie zu finden. Die Protein-Protein-Bindungsflächen sind deutlich weniger hydrophil, und die Anzahl der hydrophoben Bereiche nimmt deutlich zu. Der gleiche Trend - sogar noch etwas stärker - ist in den Protein-Ligand-Bindungsbereichen sichtbar. Die Verteilungskurve ist im hydrophilen Bereich noch etwas niedriger und die Anzahl der hydrophoben Bereiche der Proteinaußenseite erhöht. Insgesamt sind die Verteilungskurven der lokalen Lipophilie in Protein-Ligand- und Protein-Protein-Bindungsbereichen etwas breiter und flacher als in den beiden anderen Bereichen der Proteinoberfläche.

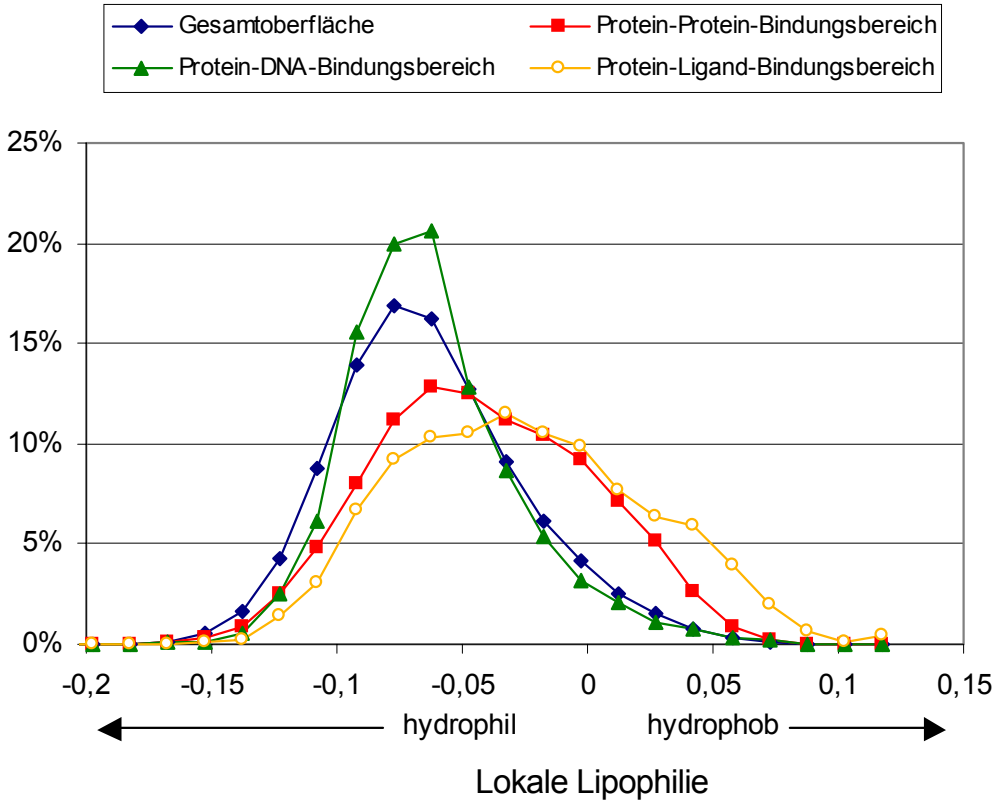


Diagramm 5.28: Lokale Lipophilie der Oberflächensegmente von Proteinen des Gesamtdatensatzes A.

5.7.3 Wasserstoffakzeptoren-/Wasserstoffdonatorendichte

In Diagramm 5.29 sind die gemittelten Dichten von Wasserstoffakzeptoren und Wasserstoffdonatoren an der Proteinoberfläche zusammengefaßt (siehe Kapitel 3.2.3.2). Die

niedrigsten Dichten sind in den Protein-Protein-Bindungsbereichen vorhanden, und die höchsten Dichten befinden sich in den Protein-DNA- und Protein-Ligand-Bindungsbereichen. Die Werte der Gesamtoberfläche liegen etwa in der Mitte dieser beiden Extrema. Wie in den vorherigen Untersuchungen des elektrostatischen Potentials und der lokalen Lipophilie liegen die Akzeptoren- und Donatorendichten in den Datensätzen A und B sehr dicht nebeneinander, weichen jedoch in den Datensätzen C und D etwas ab. An der Gesamtoberfläche und im Protein-Protein-Bindungsbereich liegen die Werte für die Datensätze C und D jeweils ein wenig über dem Gesamtdatensatz A. In den Protein-Ligand-Bindungsstellen sind die Dichten jedoch merklich unter den Werten des Gesamtdatensatzes. Die hohen Wasserstoffakzeptoren- und -donatorendichten in den Ligand- und DNA-Bindungsbereichen sind ein Hinweis darauf, daß Wasserstoffbrücken eine wichtige Rolle bei der Stabilisierung dieser Komplexe spielen.

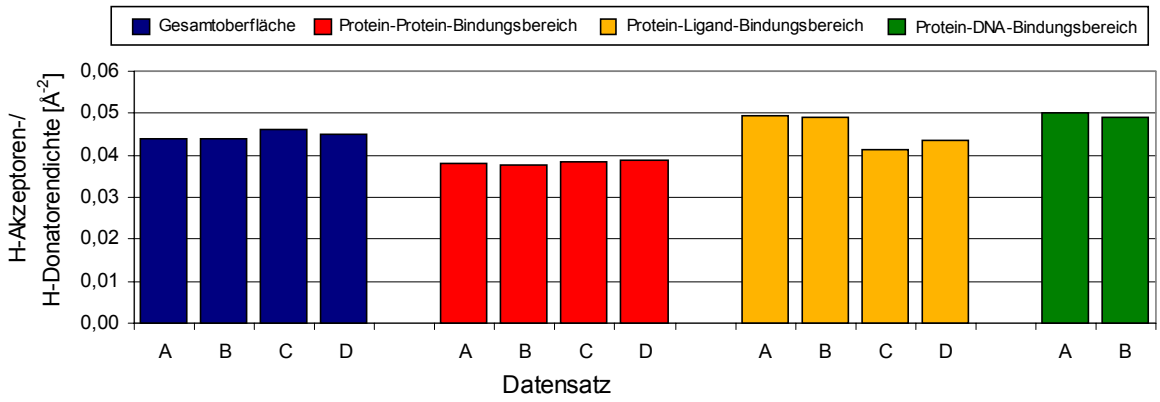


Diagramm 5.29: Wasserstoffakzeptoren-/Wasserstoffdonatorendichte gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente.

In den Verteilungskurven der Wasserstoffakzeptoren-/Wasserstoffdonatorendichte im Gesamtdatensatz A (siehe Diagramm 5.30) ist der Unterschied zwischen den verschiedenen Bindungsbereichen sehr deutlich sichtbar. Die Verteilungskurve der Dichte in den Protein-Protein-Bindungsbereichen der molekularen Oberfläche sind nach links zu niedrigerer Dichte verschoben, während in den DNA-Bindungsstellen höhere Dichten beobachtet werden. In den Ligandbindungsstellen ist die Verteilung allgemein sehr viel breiter und erstreckt sich über den gesamten Bereich zwischen 0,0 und 0,1 $Å^{-2}$. Die höchsten Dichtewerte (>0.085 $Å^{-2}$) werden in den Protein-Ligand-Bindungsbereichen erreicht.

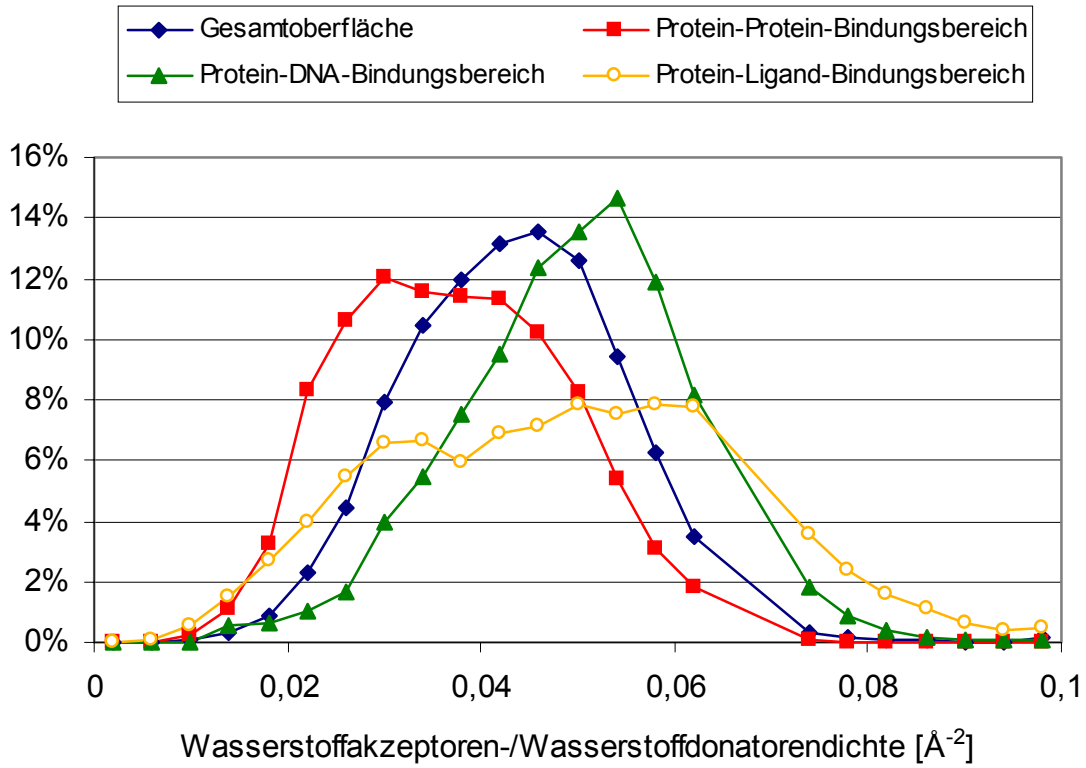


Diagramm 5.30: Wasserstoffakzeptoren-/Wasserstoffdonatorendichte der Teiloberflächen von Proteinen des Gesamtdatensatzes A.

In dem folgenden Diagramm 5.31 sind die Mittelwerte für die Akzeptoren- und Donatorendichte getrennt dargestellt. Die Mittelwerte für die Akzeptorendichte und Donatorendichte unterscheiden sich in den Protein-DNA-Bindungsbereichen drastisch. Die Wasserstoffdonatorendichte ist fast doppelt so hoch wie die Wasserstoffakzeptorendichte. Dabei ist die Akzeptorendichte der DNA-Bindungsstellen fast identisch mit den Dichte in den Protein-Protein-Bindungsstellen. Durch diese Erhöhung der Wasserstoffdonatoren-dichte können mehr Wasserstoffbrücken zu den Wasserstoffakzeptoren des Zucker-Phosphat-Rückgrats der DNA ausgebildet und somit die Protein-DNA-Komplexbindung gefestigt werden. Diese Beobachtung stimmt mit den Ergebnissen aus Kapitel 5.2.3 überein. In den anderen Oberflächenbereichen (Gesamtoberfläche, Protein-Protein- und Protein-Ligand-Bindungsregionen) unterscheiden sich die mittlere Akzeptoren- und Donatorendichte nur geringfügig. Die Akzeptorendichte ist insgesamt etwas höher als die Donatorendichte. Das Verhältnis der drei verschiedenen Bindungsbereiche zueinander entspricht den schon in Diagramm 5.29 gezeigten Verhältnissen.

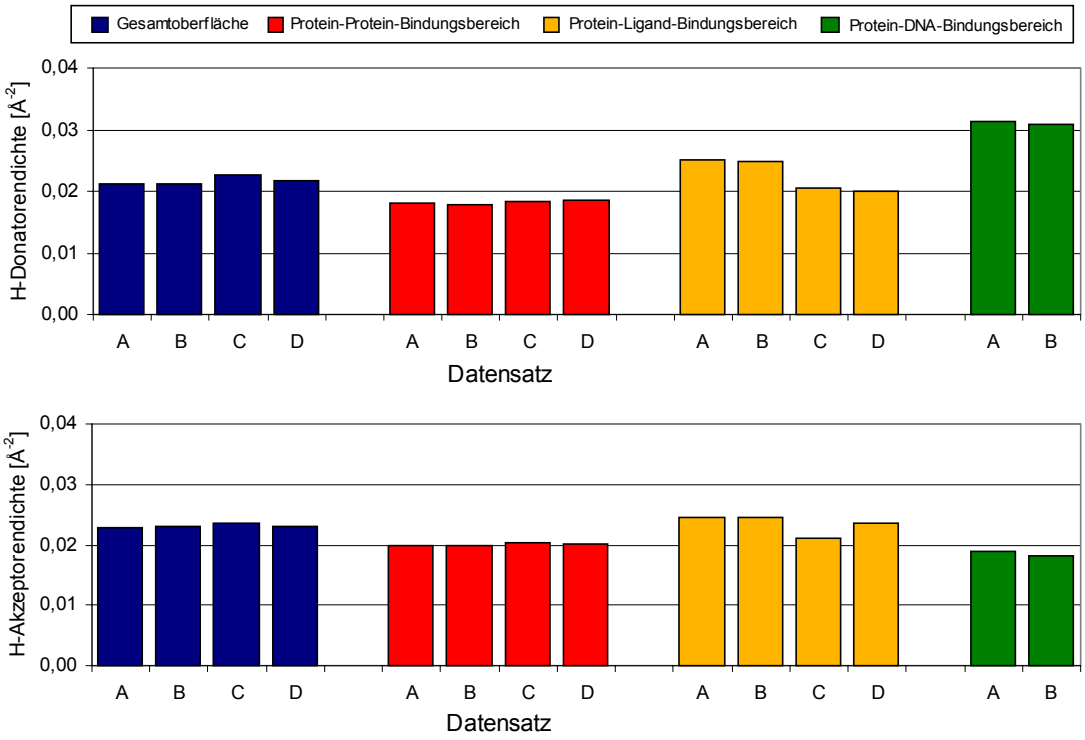


Diagramm 5.31: Wasserstoffdonatorendichte (oben) und Wasserstoffakzeptorendichte (unten) gemittelt über die Oberflächenpunkte der Teiloberflächen.

Die Verteilungskurven der Wasserstoffdonatoren- bzw. Wasserstoffakzeptorendichte im Diagramm 5.32 zeigen den vorgenannten Unterschied im DNA-Bindungsbereich der Proteinkomplexe. Die Verteilung der Wasserstoffakzeptorendichte an der DNA-Bindungsoberfläche entspricht der Verteilung in den Protein-Protein-Bindungsbereichen, d.h. die Akzeptorendichte ist niedriger als über die gesamte Proteinoberfläche betrachtet. Die Donatorendichte ist jedoch im DNA-Bindungsbereich stark erhöht, während sie für die Protein-Protein-Bindungsregionen gegenüber der Gesamtoberfläche ebenfalls niedriger ist.

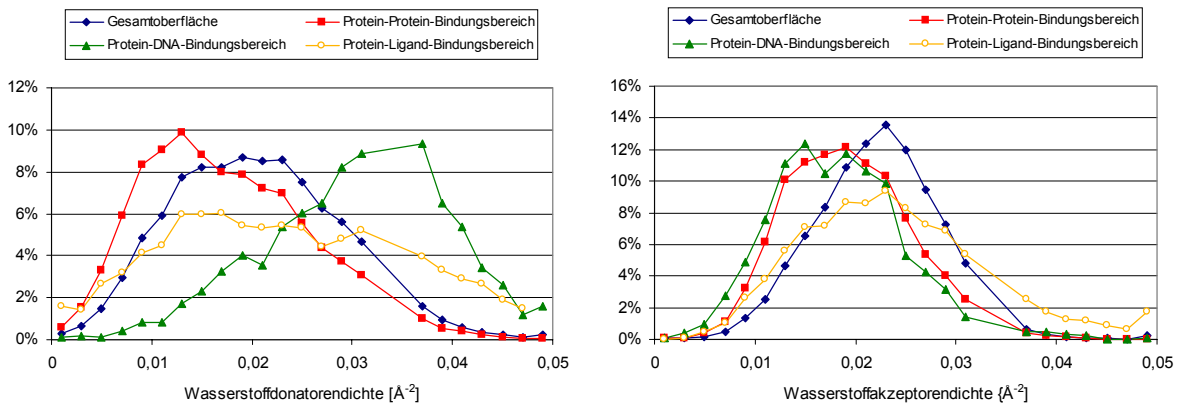


Diagramm 5.32: Wasserstoffdonatorendichte (links) und Wasserstoffakzeptorendichte (rechts) der Teiloberflächen von Proteinen des Gesamtdatensatzes A.

5.7.4 Tiefeninformation

Die Mittelwerte der Tiefeninformation (siehe Kapitel 3.2.5) der Oberflächensegmente in den verschiedenen Datensätzen und Oberflächenbereichen sind in Diagramm 5.33 zusammengefaßt. Die gemittelten Tiefen der Taschen und Spalten in den Proteinoberflächen sind an den Gesamtoberflächen und in den Protein-Protein-Bindungsbereichen gering (ca. 1,0-1,4 Å). Diese niedrigen Tiefenwerte kommen durch die Rauheit der molekularen Oberflächen zustande. In den Protein-DNA-Bindungsregionen ist der gemittelte Tiefenwert etwas größer, aber immer noch unter 2 Å. Sehr tiefe Oberflächentaschen und -spalten sind in den Ligandbindungsbereichen zu finden. Dabei fällt die Sonderstellung des Datensatzes C (Antigen-Antikörper-Komplexe) auf. Sowohl an der Gesamtoberfläche als auch in den Protein-Protein-Bindungsbereichen sind die Tiefenwerte etwas niedriger als in allen anderen Datensätzen. In den Protein-Ligand-Bindungsregionen tritt diese Erniedrigung noch drastischer hervor. Die mittlere Tiefe beträgt dort im Antigen-Antikörper-Datensatz weniger als die Hälfte der Tiefe in den anderen Datensätzen.

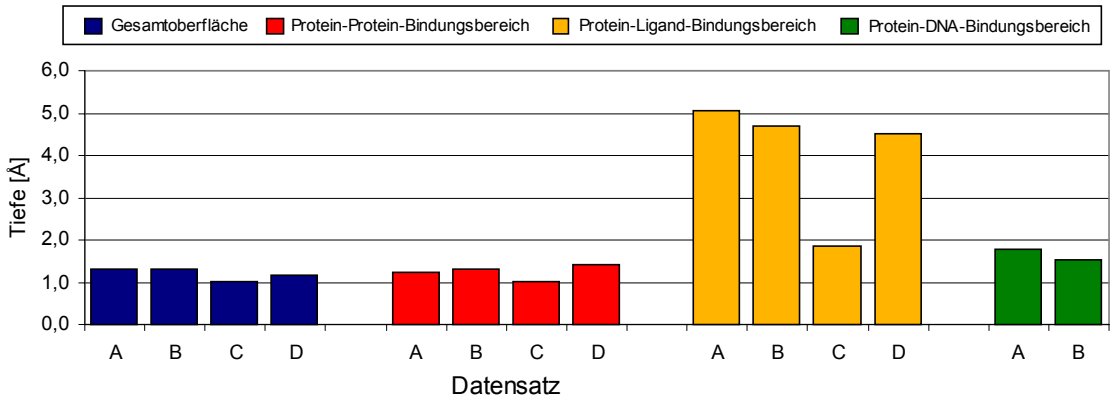


Diagramm 5.33: Tiefeninformation gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente.

Die Verteilung der mittleren Tiefen von Taschen und Spalten in den Gesamtoberflächen, Protein-Protein- und Protein-DNA-Bindungsbereichen sind sehr ähnlich (Diagramm 5.34). Der Hauptteil der Oberfläche besitzt einen Tiefenwert von unter 2 Å, welcher auf die schon angesprochene Oberflächenrauheit und die Methode zur Bestimmung der Tiefeninformation zurückzuführen ist. Nur wenige Oberflächenbereiche befinden sich in tieferen Taschen bzw. Spalten der Proteinoberfläche. In den DNA-Bindungsbereichen ist der Anteil von Oberflächensegmenten, die Taschen mit einer Tiefe bis 6 Å bilden, ein wenig größer. In den Protein-Ligand-Bindungsbereichen sieht die Verteilung anders aus. Sie ist flach und breit, ohne ausgeprägtes Maximum. Es treten Tiefenwerte in einem sehr breiten Bereich

bis weit über 10 Å auf. Aus diesen Ergebnissen kann man folgendes schließen: Protein-Protein-Bindungsbereiche unterscheiden sich in ihrer Form nicht von der Gesamtoberfläche der Proteine. DNA-Moleküle haben eine leichte Tendenz, in Spalten oder Taschen der Proteinoberfläche zu binden. Liganden binden hingegen bevorzugt in tiefen Spalten oder Höhlen der Proteinoberfläche. Im Mittel sind diese Ligandbindungstaschen ca. 4-5 Å tief. Die Antigen-Antikörper-Komplexe unterscheiden sich von den anderen Datensätzen. Sie haben weniger tiefe Spalten in der Gesamtoberfläche und den Protein-Protein-Bindungsbereichen. Auch sind Liganden nicht in tiefen Spalten oder Taschen der Oberfläche gebunden wie in den anderen Datensätzen.

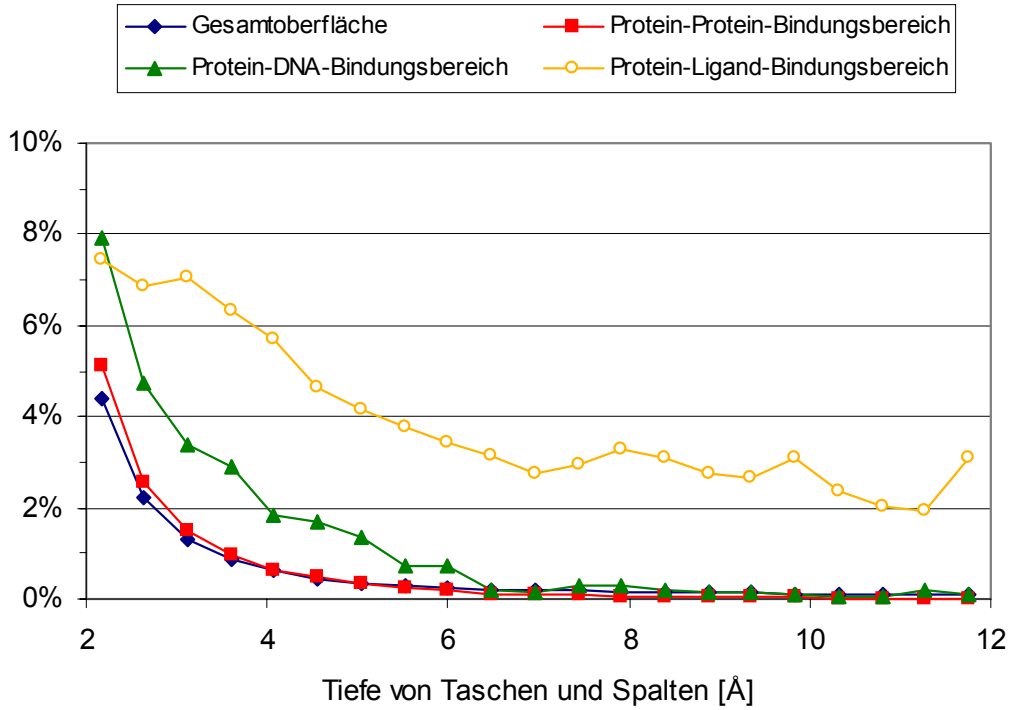


Diagramm 5.34: Tiefeninformation der Oberflächensegmente von Proteinen des Gesamtdatensatzes A.

5.7.5 Oberflächenkrümmung

In Diagramm 5.35 werden die mittleren Krümmungen (STI-Werte) der Teiloberflächen in den verschiedenen Bereichen der Proteinoberflächen miteinander verglichen (siehe Kapitel 3.2.4). Sie unterscheiden sich aufgrund der Mittelung nur geringfügig. Es lassen sich nur noch Tendenzen der Oberflächenteilbereiche zu konvexen oder konkaven Formen erkennen. Werte kleiner 2 beschreiben konkave Formen (Spalt oder Loch), Werte größer 2 repräsentieren konvexe Formen (Grat oder Pfropf).

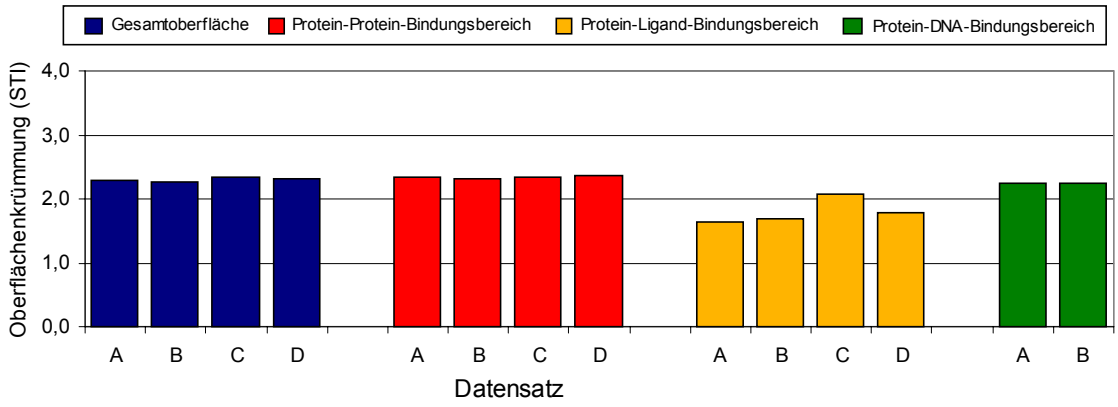


Diagramm 5.35: Oberflächenkrümmung gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente.

Der Mittelwert der STI-Werte der Oberflächensegmente an der Proteingesamtoberfläche liegt bei etwa 2,3. Das weist auf sattelförmige Oberflächensegmente mit einer leichten Tendenz zur Konvexität hin. Die konvexe Grundtendenz ist durch die globuläre Gestalt der Proteine bedingt. In den Protein-Protein-Bindungsbereichen betragen die STI-Mittelwerte ebenfalls ca. 2,3. Die STI-Werte der Protein-DNA-Bindungsbereiche sind geringfügig niedriger (ca. 2,2). Sehr viel geringer sind die STI-Werte jedoch in den Ligandbindungsstellen der Proteinoberflächen (mit Ausnahme der Antigen-Antikörper-Komplexe). Die Ligandbindungsstellen haben also eine Tendenz zu konkaven Oberflächenbereichen. Das steht im Einklang mit den vorhergehenden Ergebnissen (Tiefeninformation), daß Ligandbindungsstellen sich in Spalten und Höhlen der Oberfläche befinden. Auch die Ausnahme des STI-Mittelwertes der Protein-Ligand-Bindungsbereiche der Proteine in Datensatz C stimmt mit den vorgenannten Ergebnissen der Tiefeninformation überein. Ligandbindungsbereiche in Antigen-Antikörper-Komplexen sind nur wenig in den Proteinen vergraben (mittlere Oberflächentiefe = 1,85 Å) und im Vergleich zu den Ligandbindungsstellen von Proteinen aus den anderen Datensätzen nicht besonders konkav (mittlerer STI = 2,06).

Die Verteilung der Oberflächenkrümmung der Proteinteiloberflächen ist in Diagramm 5.36 dargestellt. Die Verteilung der STI-Werte an der Gesamtoberfläche und in den Protein-Protein- bzw. Protein-DNA-Bindungsbereichen ist fast identisch. In den Protein-DNA-Bindungsbereichen sind etwas mehr Segmente mit einem niedrigeren STI-Wert anzutreffen, d.h. die DNA-Bindungsbereiche von Proteinen haben im Mittel eine kleine Tendenz zu mehr konkaven Oberflächenregionen. Die Oberflächensegmente in den Protein-Ligand-Bindungsbereichen haben deutlich niedrigere STI-Werte. Die gesamte Verteilungskurve ist in den konkaven Bereich der Oberflächenkrümmung verschoben.

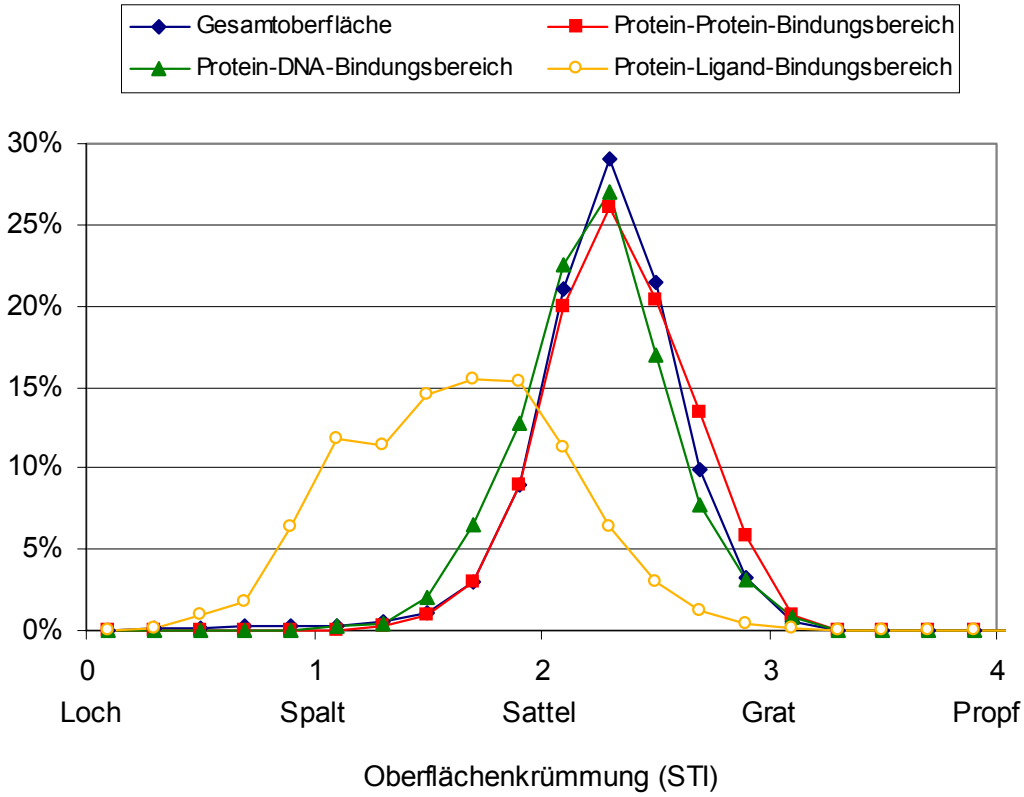


Diagramm 5.36: Oberflächenkrümmung der Oberflächensegmente von Proteinen des Gesamtdatensatzes A.

5.7.6 Flexibilität

Die lokale Flexibilität (Debye-Waller-Temperaturfaktoren – siehe Kapitel 3.2.6) der Atome an der Gesamtoberfläche, die zum größten Teil aus nichtbindenden Oberflächenbereichen besteht, ist höher als in den Bindungsbereichen der Proteinoberfläche (Diagramm 5.37 und 5.38). Durch die Bindung zu anderen Molekülen sind die Oberflächenatome in den Bindungsregionen stärker fixiert und in ihrer Bewegungsfreiheit eingeschränkt. Die niedrigste Flexibilität besitzen Atome in den Ligandbindungsregionen. Diese Bereiche befinden sich teilweise tief im Inneren von Taschen und Höhlen der Proteine und sind dadurch weniger beweglich (siehe Kapitel 5.7.4). Die Protein-Protein-Bindungsbereiche sind etwas flexibler, und die höchste Flexibilität besitzen die DNA-Bindungsregionen, die fast dem Zahlenwert der Gesamtoberfläche entspricht. Zwischen den einzelnen Datensätzen gibt es Unterschiede. Der Enzym-Inhibitor-Datensatz D weicht am meisten von den anderen Datensätzen ab. An der Gesamtoberfläche und in den Protein-Protein-Bindungsbereichen liegen höhere Flexibilitäten als in den anderen drei Datensätzen vor. Im Datensatz C (Antigen-Antikörper-Komplexe) ist die Flexibilität der Ligandbindungsbereiche gegenüber den anderen Datensätzen erniedrigt. Dieses Ergebnis

überrascht ein wenig, da in diesem Datensatz die Liganden laut den bisher diskutierten Ergebnissen nicht in tiefen Spalten der molekularen Oberfläche binden.

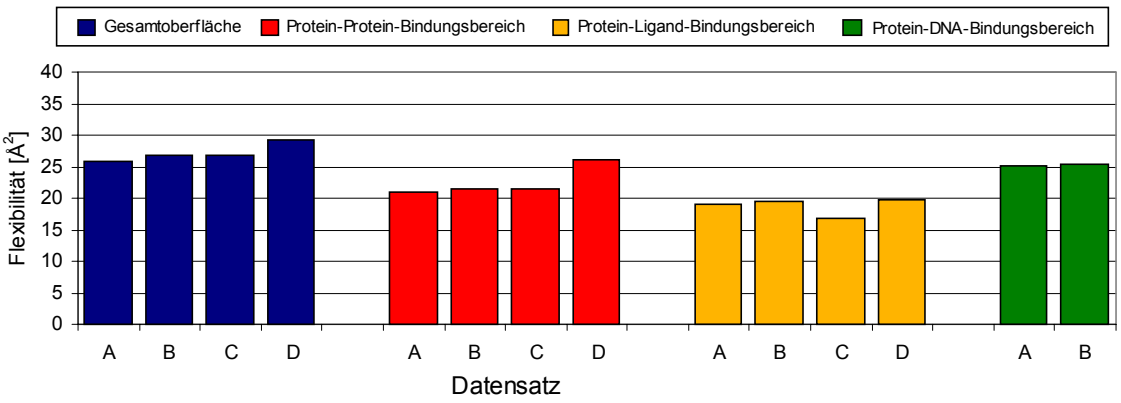


Diagramm 5.37: Flexibilität gemittelt über die Oberflächenpunkte der Oberflächenteilsegmente.

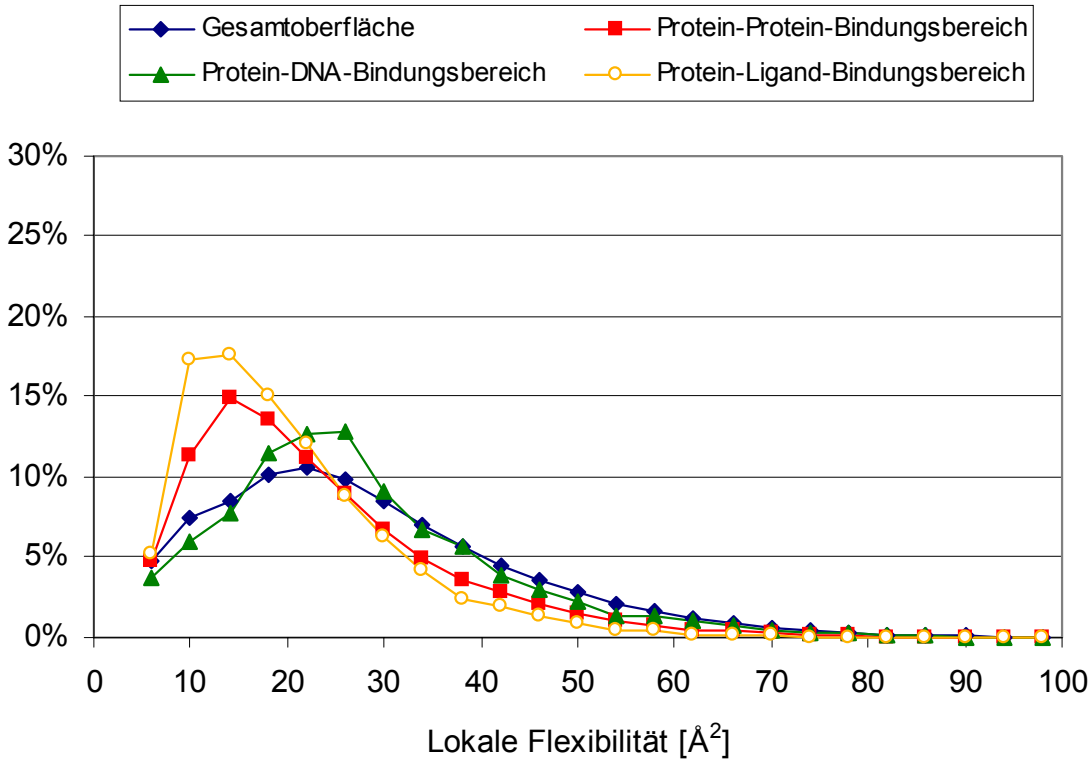


Diagramm 5.38: Flexibilität der Oberflächensegmente von Proteinen des Gesamtdatensatzes A.

5.7.7 Unterschiede der Eigenschaften in verschiedenen Oberflächenbereichen

Tabelle 5.14 faßt nochmals die wichtigsten Ergebnisse der vorangegangenen Untersuchung der sieben molekularen Eigenschaften zusammen:

Tab. 5.14: Zusammenfassung der molekularen Eigenschaften in den drei verschiedenen Bindungsbereichen der Proteine im Vergleich zu den auf die Gesamtoberfläche projizierten Eigenschaften. Die Unterschiede zwischen den molekularen Eigenschaften in den Bindungsbereichen und den Mittelwerten an der Gesamtoberfläche sind durch fünf Symbole dargestellt:
-- Die Werte in den Bindungsregionen sind eindeutig niedriger.
- Die Werte sind etwas niedriger.
O Die Werte unterscheiden sich nicht oder nur geringfügig.
+ Die Werte sind etwas größer.
++ Die Werte sind erheblich größer.

Molekulare Eigenschaft	Protein-Protein-Bindungsregion	Protein-DNA-Bindungsregion	Protein-Ligand-Bindungsregion	Protein-Ligand-Bindungsregion (Datensatz C)
Elektrostatisches Potential	O	++	+	O
Lokale Lipophilie	++	O	++	++
Wasserstoffakzeptorendichte	-	-	+	-
Wasserstoffdonatorendichte	-	++	+	-
Tiefeninformation	O	+	++	+
Oberflächenkrümmung	O	-	--	-
Lokale Flexibilität	--	-	--	--

Die Bindungsbereiche der Proteine (Protein-, DNA- und Ligandbindungsbereiche) unterscheiden sich in ihren lokalen physikalischen und chemischen Eigenschaften von der restlichen Oberfläche (bzw. der Gesamtoberfläche) und auch untereinander. Die größte Ähnlichkeit besteht zwischen den Protein-Protein-Bindungsbereichen und der Gesamtoberfläche. Die Mittelwerte von drei Eigenschaften (elektrostatisches Potential, Tiefeninformation und Oberflächenkrümmung) stimmen in diesen beiden Bereichen weitgehend überein. Der größte Unterschied zu den Gesamtoberflächen zeigt sich in den Werten der lokalen Lipophilie. Die Protein-Protein-Bindungsstellen sind deutlich hydrophober als die Gesamtoberfläche der Proteine, sie ähneln damit mehr dem Proteininneren als der Außenseite der Moleküle. Der markanteste Unterschied zwischen den DNA-Bindungsstellen und der Gesamtoberfläche ist die höhere Wasserstoffdonatorendichte und das positive elektrostatische Potential in den DNA-Bindungsbereichen. Die Ligandbindungsstellen zeichnen sich hingegen durch ihre Lage in tiefen Taschen der Proteinoberfläche aus. Die Ligandbindungsbereiche liegen teilweise tief im hydrophoben Inneren der

Proteine, was sich auch in den Werten der Lipophilie zeigt, sie sind deutlich hydrophober als die Proteinaußenseite.

Zwischen den vier verschiedenen Datensätzen fallen die Unterschiede gering aus. Die Daten in Kapitel 5.4 (Aminosäurenzusammensetzung auf der Oberfläche) hätten vermuten lassen, daß sich auch die molekularen Eigenschaften aufgrund der teilweise stark unterschiedlichen Zusammensetzung der Proteinoberfläche unterscheiden. Eine Ausnahme sind hierbei die Ligandbindungsstellen des Antigen-Antikörper-Datensatzes. Deren Eigenschaften weichen von den Ligandbindungsstellen in den anderen Datensätzen signifikant ab, deswegen werden sie in Tabelle 5.14 gesondert aufgeführt.

Die Unterschiede der physikochemischen Eigenschaften in den verschiedenen Bindungsbereichen und an der Gesamtoberfläche lassen den Schluß zu, daß es möglich ist, einen Algorithmus zu formulieren, der anhand der lokalen Eigenschaften vorhersagen kann, ob die betrachtete Teiloberfläche eine potentielle Protein-, DNA- oder Ligandbindungsstelle ist. Dabei können alle untersuchten Eigenschaften bis auf die molekulare Flexibilität in die Vorhersagemethode einfließen. Die beobachtete molekulare Flexibilität ist im Gegensatz zu den anderen Eigenschaften keine direkte Eigenschaft des einzelnen Proteinmoleküls, sondern hängt von der Struktur des Proteinkomplexes und der Strukturbestimmungsmethode ab.

6 Vorhersage von Bindungsbereichen auf Proteinoberflächen

Die Analyse der Oberflächenteilbereiche in Kapitel 5.7 zeigt, daß eine Vorhersage von Bindungsbereichen in Proteinen auf der Basis der lokalen physikalischen und chemischen Eigenschaften der Proteine möglich ist. Im Folgenden werden zwei Ansätze zur Lösung dieser Aufgabe vorgestellt.

6.1 Vorhersage von Bindungsbereichen mit einfachen Zielfunktionen

Bei der Vorhersage von Protein-Ligand-Bindungskonstanten werden oft einfache Zielfunktionen verwendet, die experimentell bestimmte Komplexbildungskonstanten mit berechneten oder gemessenen Eigenschaften der beteiligten Moleküle korrelieren [z.B. 38,133,134]. Leider sind nur für wenige Protein-Protein- und Protein-DNA-Komplexe diese Bildungskonstanten verfügbar [66]. Jedoch sind aus der vorliegenden Arbeit die lokalen Eigenschaften der Proteine und die Bindungsbereiche aller Proteinkomplexe bekannt (Kapitel 5.7). Diese Information kann für die Entwicklung eines Vorhersagealgorithmus genutzt werden. Der Aufstellung der Zielfunktion wird die Annahme zugrunde gelegt, daß der Bindungsanteil der Teiloberflächen (Kapitel 4.1.5) durch eine Linearkombination der auf die Teiloberflächen projizierten und gemittelten molekularen Eigenschaften wiedergegeben werden kann. Um auch zwischen den verschiedenen Arten von Bindungsbereichen (Protein-, DNA-, Ligandbindungsstelle und nichtbindende Oberfläche) unterscheiden zu können, wird die Zielfunktion (Gleichung 6.1) jeweils für die verschiedenen Oberflächentypen getrennt parametrisiert. Es ergeben sich somit vier verschiedene Zielfunktionen, die folgende molekulare Eigenschaften kombinieren: Elektrostatisches Potential, lokale Lipophilie, Wasserstoffdonatorendichte, Wasserstoffakzeptorendichte, Tiefeninformation und Oberflächenkrümmung.

$$X = c_0 + c_1 \cdot ESP + c_2 \cdot MLP + c_3 \cdot HA + c_4 \cdot HD + c_5 \cdot TI + c_6 \cdot STI \qquad (6.1)$$

mit

- X: Anteil der Teiloberfläche an einem Bindungsbereich bzw. nichtbindender Oberfläche
- c_0-c_6: Dimensionsbehaftete Koeffizienten der Zielfunktion
- ESP: Mittleres elektrostatisches Potential im Bereich der Teiloberfläche
- MLP: Lokale Lipophilie im Bereich der Teiloberfläche

HA: Wasserstoffakzeptorendichte im Bereich der Teiloberfläche

HD: Wasserstoffdonatorendichte im Bereich der Teiloberfläche

TI: Tiefeninformation im Bereich der Teiloberfläche

STI: Oberflächenkrümmung im Bereich der Teiloberfläche

Die Koeffizienten c_0 bis c_6 werden durch multilineare Regression mit dem Programm MATHEMATICA Version 4 (Wolfram Research, Inc.) bestimmt. Dazu stehen die in Kapitel 5.7 beschriebenen Daten der 1,2 Millionen Teiloberflächen des Gesamtdatensatzes zur Verfügung. Die Koeffizienten sowie die Korrelationskoeffizienten R^2 für die vier Funktionen sind in Tabelle 6.1 aufgelistet. Die Korrelationskoeffizienten sind sehr niedrig und widersprechen der Annahme, daß die Bindungsbereiche von Proteinen durch eine einfache Linearkombination der molekularen Eigenschaften vorhergesagt werden können.

Tabelle 6.1: Koeffizienten der Zielfunktion I (Parametrisierung mit allen Teiloberflächen des Gesamtdatensatzes).

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	R^2
Protein	0,0971	-0,0007759	1,359	-4,736	0,992	0,00631	0,0784	**0,0146**
DNA	-0,0066	0,0000878	-0,008	-0,045	0,149	0,00930	0,0026	**0,0858**
Ligand	-0,0449	-0,0000723	0,094	-0,433	0,442	0,01880	0,0160	**0,1195**
nichtbindend	0,9687	0,0007485	-1,457	5,306	-1,641	-0,03178	-0,1023	**0,0993**

Nun stellt sich die Frage, ob dies am Konzept der Zielfunktion oder an den verwendeten Daten zur Parametrisierung liegt. Um dies aufzuklären, wird die Parametrisierung nochmals mit einem kleineren Satz von Teiloberflächen wiederholt. Die nichtbindenden Teiloberflächen sind im Gesamtdatensatz sehr oft vertreten. Protein-DNA- und Protein-Ligand-Bindungsbereiche kommen jedoch nur sehr selten vor (Tabelle 5.13). Um die Zielfunktionen auf eine möglichst gute Erkennung bzw. Vorhersage aller vier Oberflächenbereiche optimieren zu können, wird in dem verkleinerten Satz von Teiloberflächen der Anteil der nichtbindenden Oberflächen reduziert und im Gegenzug der Anteil von DNA- bzw. Ligandbindungsbereichen erhöht. Insgesamt besteht der kleinere Satz aus 8161 Teiloberflächen. Er besteht aus 33% Protein-, 10% DNA-, 12% Ligandbindungsstellen und 45% nichtbindenden Teiloberflächen. Tabelle 6.2 enthält das Ergebnis der erneuten Parametrisierung. Die Korrelationskoeffizienten sind hier deutlich höher (insbesondere bei der Zielfunktion zur Vorhersage der Ligandbindungsbereiche), jedoch immer noch weit entfernt von befriedigenden Werten.

Tabelle 6.2: Koeffizienten der Zielfunktion II (Parametrisierung mit dem verkleinerten Satz von Teiloberflächen).

	c_0	c_1	c_2	c_3	c_4	c_5	c_6	$\mathbf{R^2}$
Protein	0,1401	-0,001768	1,416	-7,37	-0,290	0,00917	0,1467	**0,1301**
DNA	-0,0948	0,001041	0,160	-3,25	4,295	0,01087	0,0437	**0,1928**
Ligand	0,0572	-0,000312	0,368	-1,73	0,051	0,06594	-0,0213	**0,5389**
nichtbindend	0,8895	0,001029	-1,636	12,43	-4,166	-0,08628	-0,1666	**0,2535**

Viel interessanter als die Korrelationskoeffizienten ist die Frage, ob die Zielfunktionen in der Lage sind, die untersuchten Teiloberflächen des Gesamtdatensatzes entsprechend ihren beobachteten Bindungsverhältnissen korrekt zu klassifizieren. Dies kann mit einer Analyse der Erkennungsraten der vier Oberflächenbereiche beantwortet werden. Dazu werden für alle Teiloberflächen die Bindungsanteile X mit den parametrisierten Zielfunktionen berechnet. Ähnlich wie bei der Klassifizierung der Teiloberflächen anhand der beobachteten Bindungsverhältnisse bestimmt die Zielfunktion mit dem höchsten Bindungsanteil X die Bindungsbereichsklasse. Wenn keiner dieser Werte 20% übersteigt, wird die Teiloberfläche in die Klasse der nichtbindenden Teiloberflächen eingeordnet. Anschließend wird diese Klassifizierung mit den beobachteten Bindungsverhältnissen verglichen, d.h. es wird der Prozentsatz von Teiloberflächen bestimmt, die von dem neuronalen Netz bzw. der Zielfunktionen korrekt erkannt werden. Dieser Anteil wird im Folgendem Erkennungsrate genannt. Für jede der vier Oberflächentypen wird eine Erkennungsrate S berechnet:

$$S_\alpha = \frac{m_\alpha}{n_\alpha} \qquad \text{wobei} \quad \alpha = \{\text{Prot.}, \text{Lig.}, \text{DNA}, \text{n.bind.}\} \tag{6.2}$$

mit

S_α: Erkennungsrate der Teiloberflächen in Protein-, DNA-, Ligandbindungsregionen und nicht bindenden Oberflächenbereichen

m_α: Anzahl der vom neuronalen Netz richtig erkannten Teiloberflächen in Protein-, DNA-, Ligandbindungsregionen und nicht bindenden Oberflächenbereichen

n_α: Anzahl Teiloberflächen in Protein-, DNA-, Ligandbindungsregionen und nicht bindenden Oberflächenbereichen

Die Erkennungsraten für beide Zielfunktionen sind in Tabelle 6.3 enthalten.

Tabelle 6.3: Vorhersagegenauigkeit der vorgestellten Zielfunktionen.

Erkennungsrate:	Zielfunktion I (Parametrisierung mit Gesamtdatensatz)	Zielfunktion II (Parametrisierung mit verkleinertem Datensatz)
Protein-Protein-Bindungsstellen ($S_{Prot.}$)	0%	5,8%
Protein-DNA-Bindungsstellen (S_{DNA})	0%	2,0%
Protein-Ligand-Bindungsstellen ($S_{Lig.}$)	0%	43,8%
nichtbindende Oberflächen ($S_{n.bind.}$)	100%	97,8%

Die Zielfunktion I klassifiziert alle Teiloberflächen als nichtbindende Oberflächen und erkennt dementsprechend diese zu 100%. Zur Erkennung von Bindungsbereichen auf Proteinoberflächen ist sie jedoch völlig ungeeignet. Keine einzige Bindungsstelle wird vorhergesagt bzw. erkannt. Dieses Fehlverhalten ist, wie schon oben erwähnt, auf die Verwendung des Gesamtdatensatzes zur Parametrisierung zurückzuführen (Dominanz von nichtbindenden Teiloberflächen). In der zweiten Zielfunktion werden die Protein-Protein- und Protein-DNA-Bindungsstellen zu 6% bzw. 2% richtig erkannt. Ligandbindungsstellen werden sogar zu fast 44% korrekt vorhergesagt. Dies ist durch die besondere Lage der Ligandbindungsbereiche auf der molekularen Oberfläche bedingt. Die Liganden binden in tiefen Taschen und Höhlen (Kapitel 5.7.4), was diese Bindungsbereiche von den anderen Oberflächenteilen deutlich unterscheidet. Diese Zielfunktion kann zur Vorhersage von Ligandbindungsstellen benutzt werden, jedoch ist sie nicht für die Bestimmung von Protein-Protein- oder Protein-DNA-Bindungsstellen geeignet.

6.2 Vorhersage von Bindungsbereichen mittels neuronaler Netze

Wie im vorhergehenden Abschnitt gezeigt wurde, können mit den vorgestellten Zielfunktionen Bindungsbereiche von Proteinen nicht vorhergesagt werden. Im Folgenden wird deshalb ein neuronales Netzwerk beschrieben, mit dem diese Aufgabe bewältigt werden kann. Durch die Verwendung des neuronalen Netzes kann die Vorhersagequalität entscheidend verbessert werden. Zum Aufbau und Training dieses neuronalen Netzes wird der Stuttgarter Neuronale Netze Simulator (SNNS Version 4.2) benutzt [135].

6.2.1 Theoretische Grundlagen neuronaler Netze

Klassische Computeralgorithmen sind bei komplexen Mustererkennungsproblemen, wie z.B. der Erkennung von Gesichtern, dem menschlichen Gehirn weit unterlegen. Die Idee neuronaler Netze ist daher, die Arbeitsweise des Gehirns in Algorithmen zu übertragen, um solche Probleme mit dem Computer bearbeiten zu können.

Im Gehirn wird die Informationsverarbeitung durch sehr viele Nervenzellen bewerkstelligt. Die Nervenzellen sind im Verhältnis zum Gesamtsystem sehr einfach aufgebaut und leiten den Grad ihrer Erregung über Nervenfasern an andere Nervenzellen weiter. Künstliche neuronale Netze sind stark vereinfachte Modelle der Nervensysteme von Säugetieren. Sie bestehen in Analogie zu den biologischen Systemen aus einer großen Anzahl einfacher Einheiten (Zellen, Neuronen, Elemente, *Units*), die sich Informationen in Form der Aktivierung der Zellen über gerichtete Verbindungen zusenden. Neuronale Netze sind massiv parallele, lernfähige Systeme, mit denen komplexe Aufgaben in vielen Anwendungsgebieten gelöst werden können. Eine herausragende Eigenschaft neuronaler Netze ist ihre Lernfähigkeit, also die Fähigkeit, eine Aufgabe, wie etwa ein Klassifikationsproblem, selbständig aus Trainingsbeispielen zu lernen, ohne daß sie dazu explizit programmiert werden müssen [10,136].

6.2.1.1 Aufbau neuronaler Netze

Je nach Aufgabenstellung werden unterschiedliche, teilweise explizit angepaßte Netzwerktypen verwendet. Für die in dieser Arbeit gestellten Probleme ist die Verwendung eines einfachen Netzes ohne Rückkopplung (*Feedforward*-Netz) ausreichend. In diesem Netzwerktyp sind die einzelnen Zellen nur in einer Richtung miteinander verbunden. Im Folgenden werden neben den allgemeingültigen Regeln und Bestandteilen von neuronalen Netzen die spezifischen Details nur dieses Netzwerktypus beschrieben [136].

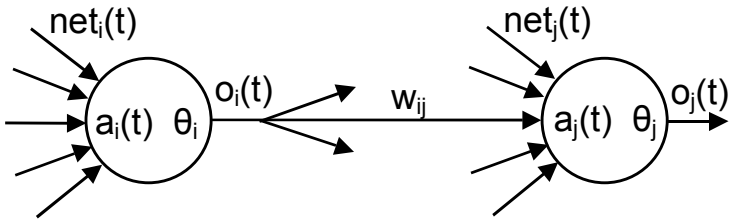


Abbildung 6.1: Darstellung der Zellen und deren Verbindung in einem neuronalen Netz

Die Zellen der neuronalen Netze sind gegenüber dem biologischen Vorbild stark vereinfacht und können nur einen Teil der biologischen Eigenschaften modellieren. Sie sind über Ein- und Ausgänge mit den anderen Zellen des Netzwerkes verbunden (siehe Abbildung 6.1). Jede Zelle ist durch ihren Aktvierungszustand (*Activation*) $a_i(t)$ am Simulationsschritt t charakterisiert. Mit der Aktivierungsfunktion f_{act} wird während jedes Lernzyklus der neue Aktivierungszustand $a_i(t+1)$ des Neurons i aus der alten Aktivierung $a_i(t)$, der Netzeingabe $net_i(t)$ und dem Schwellenwert θ_i des Neurons i berechnet:

$$a_i(t+1) = f_{act}(a_i(t), net_i(t), \theta_i) \tag{6.3}$$

Hierbei wird häufig die logistische Aktivierungsfunktion verwendet.

$$f_{act}(x) = \frac{1}{1+e^{-x}} \tag{6.4}$$

Die Ausgabe o_i der Zelle i wird über die Ausgabefunktion f_{out} aus der Aktivierung der Zelle bestimmt:

$$o_i = f_{out}(a_i) \tag{6.5}$$

In vielen neuronalen Netzen wird die Identität als Ausgabefunktion verwendet.

$$o_i = a_i = f_{out}(x) \tag{6.6}$$

Die Zellen sind wie oben erläutert über ein Verbindungsnetzwerk miteinander verknüpft. Die Zelleneingänge sind mit den Ausgängen der Vorgängerzellen verbunden. Jede Verbindung ist mit einem Gewicht versehen, welches die Signalweiterleitung zwischen den Zellen beeinflußt. Die Netzeingabe $net_j(t)$ von Neuron j berechnet sich aus der Summe der Ausgaben $o_i(t)$ der Vorgängerzellen i und den Verbindungsgewichten w_{ij} zwischen Zelle i und j (siehe Abbildung 6.1).

$$net_j(t) = \sum_i w_{ij} \cdot o_i(t) \tag{6.7}$$

Feedforward-Netze haben Verbindungen nur in eine Richtung (siehe Abbildung 6.2) [136,137]. Die Zellen werden in Schichten gruppiert und nach der Position im Netzwerk klassifiziert. Man unterscheidet zwischen einer Eingabeschicht, einer Ausgabeschicht und einer oder mehreren verdeckten Schichten. Die Zellen der Eingabeschicht nehmen die zu verarbeitende Eingabeinformation auf und leiten sie in das Netz weiter. Zellen innerhalb der Eingabeschicht werden daher als Eingabeeinheiten bezeichnet. Die Verarbeitung und Kombination der Informationen finden in den Zellen der dazwischen liegenden Schichten statt. Die Zustände dieser Zellen sind nach außen nicht sichtbar. Sie werden deshalb verdeckte Neuronen genannt. Die Anzahl verdeckter Schichten ist variabel und hängt von der Problemstellung ab.

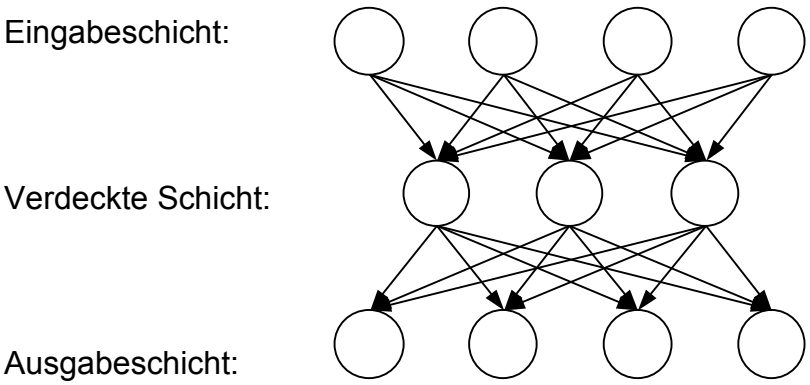


Abbildung 6.2: Darstellung eines *Feedforward*-Netzes.

Über eine Lernregel wird festgelegt, wie ein neuronales Netz lernt, vorgegebene Eingabemuster zu verarbeiten und korrekt zu klassifizieren. Das Lernverfahren besteht in der Modifikation von Netzwerkparametern, meist der Stärke der Verbindungen oder der Parameter der Aktivierungsfunktionen. Während des Lernzyklus werden dem Netzwerk wiederholt Trainingsmuster präsentiert und die Ausgabe des Netzwerkes mit der erwarteten Ausgabe verglichen. Die Lernregel verändert dann basierend auf der Differenz der erwarteten und der tatsächlichen Ausgabe die Gewichte des Verbindungsnetzwerkes. Der Vorgang wird so lang wiederholt, bis der Fehler für alle Trainingsmuster ein gewünschtes Minimum erreicht hat. Es gibt eine Vielzahl verschiedener Lernverfahren, von denen die wichtigsten im Folgenden vorgestellt werden.

6.2.1.2 Hebbsche Lernregel

Die meisten neuronalen Lernverfahren basieren auf der Hebbschen Lernregel. Sie wurde 1949 vom Physiologen Donald Hebb zur Begründung experimenteller Ergebnisse psychologischer Experimente formuliert [138]:

> "Wenn ein Axon der Zelle A nahe genug ist, um eine Zelle B zu erregen, wiederholt oder dauerhaft sich am Feuern beteiligt, geschieht ein Wachstumsprozeß oder metabolische Änderung in einer oder beiden Zellen dergestalt, daß A's Effizienz als eine der auf B ... feuernden Zellen anwächst."

Der mathematische Ausdruck dieser Hebbschen Regel lautet (nach [139]):

> „Wenn Zelle j eine Eingabe von Zelle i erhält und beide gleichzeitig stark aktiviert sind, dann erhöhe das Gewicht w_{ij} (die Stärke der Verbindung von i nach j)."

$$\Delta w_{ij} = \eta \cdot o_i \cdot a_j \tag{6.8}$$

mit

Δw_{ij}: Änderung des Gewichtes w_{ij}
η: Lernrate
o_i: Ausgabe der Vorgängerzelle i
a_j: Aktivierung der Nachfolgerzelle j

Die Hebbsche Regel läßt sich auch in einer allgemeineren Form schreiben. Die Gewichtsänderung Δw_{ij} ist das Produkt zweier Funktionen $h(o_i, w_{ij})$ und $g(a_j, a_{j,erwartet})$, die von der Ausgabe o_i der Vorgängerzelle und dem Verbindungsgewicht w_{ij} zur Zelle j, bzw. von der Aktivierung a_j der Zelle und der erwarteten Aktivierung $a_{j,erwartet}$ (*Teaching Input*) abhängen [136]:

$$\Delta w_{ij} = \eta \cdot h(o_i, w_{ij}) \cdot g(a_j, a_{j,erwartet}) \tag{6.9}$$

Diese allgemeine Formulierung ist die Basis für eine Vielzahl komplexerer Lernregeln.

6.2.1.3 Delta-Regel

Die Delta-Regel (auch als Widrow-Hoff-Regel bekannt [139]) wird aus der allgemeinen Form der Hebbschen Regel abgeleitet. Die Änderung des Verbindungsgewichts von Zelle i zu Zelle j ist proportional zur Differenz δ_j der aktuellen Aktivierung $a_j = o_j$ und der erwarteten Aktivierung $a_{j,erwartet}$.

$$\Delta w_{ij} = \eta \cdot o_i \cdot \left(a_{j,erwartet} - o_j\right) = \eta \cdot o_i \cdot \delta_j \tag{6.10}$$

6.2.1.4 Backpropagation-Regel

Backpropagation (Rückwärtspropagierung) ist eine Verallgemeinerung der oben beschriebenen Delta-Regel. Sie wird für Netze mit mehr als einer Schicht trainierbarer Gewichte und für Neuronen mit einer nichtlinearen Aktivierungsfunktion (z.B. der logistischen Funktion 6.3) verwendet. Die Gewichtsänderung ist ebenfalls von der Differenz der aktuellen und erwarteten Aktivierung abhängig. Jedoch wird die Differenz δ_j, je nachdem ob sich Zelle j in der Ausgabeschicht oder in einer verdeckten Schicht befindet, unterschiedlich berechnet:

$$\Delta w_{ij} = \eta \cdot o_i \cdot \delta_j \tag{6.11}$$

mit

$$\delta_j = \begin{cases} o_j \cdot \left(1 - o_j\right) \cdot \left(a_{j,erwartet} - o_j\right) & \text{falls j eine Ausgabezelle ist} \\ o_j \cdot \left(1 - o_j\right) \cdot \sum_k \left(\delta_k \cdot w_{jk}\right) & \text{falls j eine verdeckte Zelle ist} \end{cases}$$

Die *Backpropagation*-Regel kann durch eine einfache Methode verbessert werden. Diese Methode beruht auf der Einführung des sogenannten *Momentum*-Terms und wird auch als konjugierter Gradientenabstieg bezeichnet. Durch den zusätzlichen Term wird die bereits vorgenommene Änderung des Gewichtes Δw_{ij} am Lernschritt t beim nächsten Schritt t+1 berücksichtigt:

$$\Delta w_{ij}(t+1) = \eta \cdot o_i \cdot \delta_j + \alpha \cdot \Delta w_{ij}(t) \tag{6.12}$$

mit

t: Lernschritt t innerhalb des Lernverfahrens

α: *Momentum*-Faktor

6.2.1.5 Ablauf einer Simulation eines neuronalen Netzes

Der typische Ablauf eines überwachten Lernverfahrens, wie z.B. mit der *Backpropagation*-Regel, kann in fünf Hauptschritte unterteilt werden [136]:

1. Präsentation des Eingabemusters durch entsprechende Aktivierung der Eingabeeinheiten
2. Vorwärtspropagierung der angelegten Eingabe durch das Netz; dies erzeugt ein Ausgabemuster für die aktuelle Eingabe
3. Vergleich der Ausgabe mit der erwünschten Ausgabe (*Teaching Input*) liefert einen Fehlervektor (Differenz δ)
4. Rückwärtspropagierung der Fehler von der Ausgabeschicht zur Eingabe liefert Änderungen der Verbindungsgewichte, die dazu dienen, den Fehlervektor zu verringern
5. Änderung der Gewichte aller Neuronen des Netzes um die vorher berechneten Werte.

6.2.2 Struktur des neuronalen Netzes und Vorbereitung der Eingabemuster

Ziel des neuronalen Netzes ist es, basierend auf den molekularen Eigenschaften eines beliebigen Teilbereiches eines Proteins (repräsentiert durch eine Teiloberfläche) vorherzusagen, ob dieser Teilbereich eine mögliche Protein-, DNA- oder Ligandbindungsstelle ist. Dazu werden folgende molekulare Eigenschaften mit dem neuronalen Netz verarbeitet: Elektrostatisches Potential, lokale Lipophilie, Wasserstoffakzeptorendichte, Wasserstoffdonatorendichte, Tiefeninformation und Oberflächenkrümmung (STI). Das Netzwerk wird so aufgebaut, daß die ausgewählten Eigenschaften jeweils durch zwei oder drei Eingabeeinheiten verarbeitet werden. Je mehr Neuronen für die einzelnen Eigenschaften benutzt werden, desto genauer können die Zusammenhänge zwischen den molekularen Eigenschaften und den Bindungsbereichen durch das neuronale Netz abgebildet werden. Mit steigender Anzahl der Neuronen des Netzes nimmt jedoch die benötigte Rechenzeit für das Training des neuronalen Netzes zu. Es muß also ein Kompromiß zwischen Genauigkeit und Rechenzeit erzielt werden.

In dieser Arbeit werden zwei leicht unterschiedliche neuronale Netze erstellt. Sie unterscheiden sich in der Anzahl der Eingabeeinheiten für das elektrostatische Potential und die lokale Lipophilie. Das erste Netz (I) benutzt drei Eingabeeinheiten für das elektrostatische Potential und die Oberflächenkrümmung und zwei Neuronen für die lokale Lipophilie. Die restlichen Eigenschaften werden mit jeweils zwei Eingabeeinheiten verarbeitet. So wird je

nach Zahlenwert des elektrostatischen Potentials das Neuron für überwiegend elektronegative, neutrale oder elektropositive Teiloberflächen aktiviert. Die lokale Lipophilie wird durch ein lipophiles und ein hydrophiles Neuron repräsentiert. Im zweiten Netz (II) wird das elektrostatische Potential der Teiloberfläche in einen elektropositiven und einen elektronegativen Anteil aufgeteilt. Anstatt den Mittelwert des elektrostatischen Potentials durch Summation der Potentialwerte aller Oberflächenpunkte der Teiloberfläche zu bestimmen, wird in diesem Fall ein Mittelwert aller elektronegativen und ein Mittelwert aller elektropositiven Oberflächenpunkte gebildet. Ohne diese Aufteilung ist es bei Teiloberflächen mit einem Mittelwert nahe null nicht möglich zu unterscheiden, ob alle Oberflächenpunkte überwiegend neutral sind oder ob sich zwei gleich große elektropositive und elektronegative Teilbereiche im Mittelwert aufheben. Die beiden elektrostatischen Mittelwerte verwenden jeweils zwei Eingabeeinheiten, d.h. das elektrostatische Potential wird durch insgesamt vier Eingabeeinheiten verarbeitet (stark positive, schwach positive, schwach negative und stark negative Teiloberfläche). Für die lokale Lipophilie werden ebenfalls vier Eingabeeinheiten verwendet: Zwei Neuronen repräsentieren den lipophilen und zwei weitere den hydrophilen Teilbereich. Insgesamt besitzt Netz I 14 und Netz II 17 Eingabeeinheiten. Die Aktivierung der Eingabeeinheiten wird aus den Mittelwerten der molekularen Eigenschaften und gegebenen Minimal- und Maximalwerten berechnet. Diese Minimal- und Maximalwerte sind neben der Anzahl der Eingabeeinheiten für jede untersuchte Eigenschaft in Tabelle 6.4 und 6.5 zusammengefaßt. Die verwendeten Funktionen für die Verteilung des Eigenschaftswertes auf zwei bzw. drei Eingabeeinheiten sind in Diagramm 6.1 dargestellt. Durch die gegebenen Aktivierungsfunktionen besteht das Eingabemuster aus kontinuierlichen Aktivierungswerten im Bereich zwischen 0,0 und 1,0.

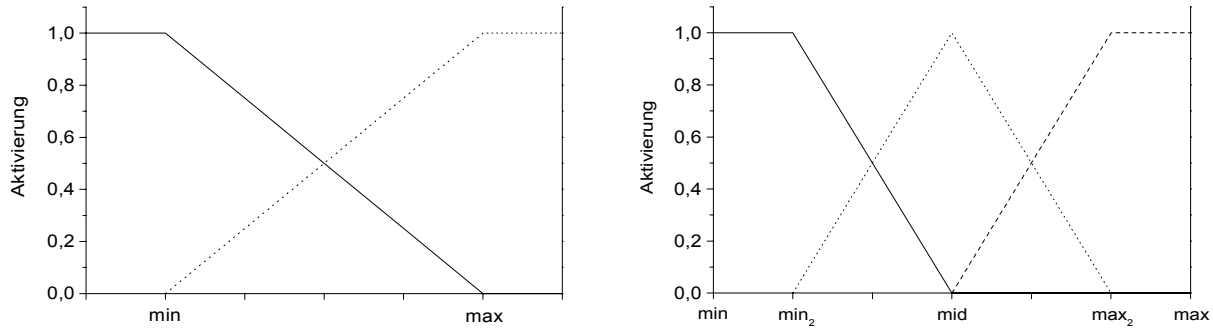


Diagramm 6.1: Aktivierungsfunktionen der Eingabeeinheiten:
Links: Eigenschaft wird durch zwei Eingabeeinheiten repräsentiert
(—) Aktivierung von Neuron 1 (···) Aktivierung von Neuron 2
Rechts: Eigenschaft wird durch drei Eingabeeinheiten repräsentiert
(—) Neuron 1 (···) Neuron 2 (--) Neuron3.

Tabelle 6.4: Parameter der Eingabeschicht des neuronalen Netzwerkes I.

Molekulare Eigenschaft	Eingabeeinheiten	min	max	min_2	mid	max_2
Elektrostatisches Potential	3	-80,00	80,00	-53,33	0,00	53,33
Lokale Lipophilie	2	-0,10	0,10	-	-	-
Wasserstoffdonatorendichte	2	0,00	0,06	-	-	-
Wasserstoffakzeptorendichte	2	0,00	0,06	-	-	-
Tiefeninformation	2	0,00	6,00	-	-	-
Oberflächenkrümmung (STI)	3	0,50	3,50	1,00	2,00	3,00

Tabelle 6.5: Parameter der Eingabeschicht des neuronalen Netzwerkes II.

Molekulare Eigenschaft	Eingabeeinheiten	min	max	min_2	mid	max_2
Elektrostatisches Potential (positiver Teilbereich)	2	0,00	100,00	-	-	-
Elektrostatisches Potential (negativer Teilbereich)	2	-100,00	0,00	-	-	-
Lokale Lipophilie (lipophiler Teilbereich)	2	0,00	0,10	-	-	-
Lokale Lipophilie (hydrophiler Teilbereich)	2	-0,25	0,00	-	-	-
Wasserstoffdonatorendichte	2	0,00	0,06	-	-	-
Wasserstoffakzeptorendichte	2	0,00	0,06	-	-	-
Tiefeninformation	2	0,00	6,00	-	-	-
Oberflächenkrümmung (STI)	3	0,50	3,50	1,00	2,00	3,00

Beide Netze verwenden eine verdeckte Schicht. Sie besteht im ersten Netzwerk aus 12 und im zweiten Netzwerk aus 13 verdeckten Neuronen. Die Ausgabeschicht ist bei beiden Netzwerken gleich und wird durch vier Ausgabeneuronen gebildet, welche die verschiedenen Oberflächenbereiche repräsentieren: Protein-Protein-, Protein-Ligand-, Protein-DNA-Bindungsbereich und nichtbindende Teiloberfläche. In beiden Netzen ist die Eingabeschicht vollständig mit der verdeckten Schicht und die verdeckte Schicht vollständig mit der Ausgabeschicht verbunden. Netzwerk I besitzt 216 und Netzwerk II 273 gewichtete Verbindungen zwischen den Neuronen. Die Struktur von Netz I ist in Abbildung 6.3 dargestellt. Als Aktivierungsfunktion für die Neuronen der versteckten Schicht und der Ausgabeschicht wird eine sigmoidale Funktion, die sogenannte logistische Aktivierungsfunktion (Gleichung 6.4), verwendet. Die Ausgabefunktion ist bei allen Neuronen die Identitätsfunktion (Gleichung 6.6). Die Aktivierung der Ausgabeneuronen gibt an, ob die zum Eingabemuster gehörende Teiloberfläche eine mögliche Bindungsstelle ist und um welche Art Bindungsstelle es sich dabei handelt. Der Bindungsstellentyp wird dabei durch das Ausgabeneuron mit der höchsten Aktivierung bestimmt (vergleiche

Kapitel 4.1.5). Wenn keines der vier Neuronen eine Aktivierung über 20% besitzt, wird die untersuchte Teiloberfläche in keine der vier Klassen eingeteilt, sondern als „nicht erkannt“ klassifiziert.

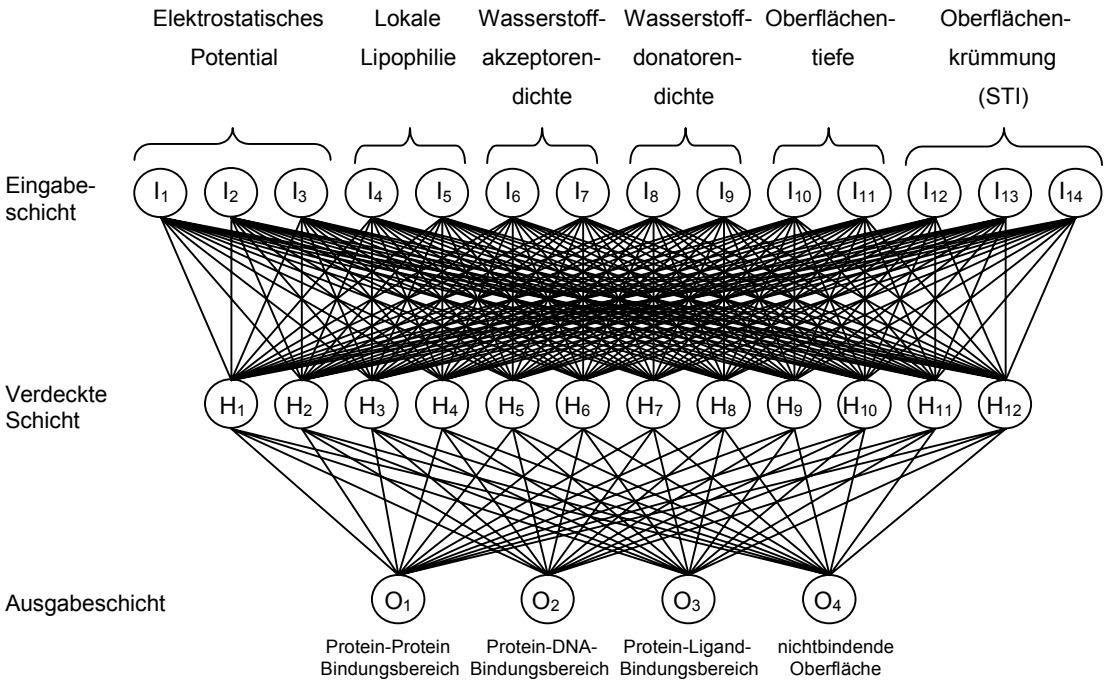


Abbildung 6.3: Neuronales Netzwerk I mit 2 Schichten (3 Zellschichten) zur Vorhersage von Bindungsbereichen.

6.2.3 Training des neuronalen Netzes

Für jedes Eingabemuster der molekularen Eigenschaften einer Teiloberfläche repräsentiert der im Proteinkomplex beobachtete Bindungsanteil der Teiloberfläche das jeweilige „korrekte“ Ausgabemuster (*Teaching Input*). Für das Training des neuronalen Netzes wird ein überwachtes Lernverfahren verwendet, d.h. für jedes Eingabemuster wird die Netzausgabe mit dem „korrekten“ Ausgabemuster verglichen. Als Lernregel wird die *Backpropagation*-Regel mit *Momentum*-Term (siehe Gleichung 6.12) verwendet. Die Parameter der *Backpropagation*-Regel (Lernrate und *Momentum*-Faktor) werden ausgehend von den SNNS-Standardwerten durch mehrere Testläufe optimiert. So wird eine Lernrate η von 0,01 und ein *Momentum*-Faktor α von 0,01 ermittelt. Die maximale Fehlerrate $a_{i,erwartet}$-o_i (vergleiche Kapitel 6.2.1.4), die nicht zu einer Gewichtsänderung führt, beträgt 0,05. Die Initialisierung der Gewichte des Netzes geschieht mit zufälligen Werten zwischen -1,0 und +1,0. Während des Trainings werden die Eingabemuster in

jedem Schritt in einer neuen zufälligen Reihenfolge abgearbeitet. Dadurch wird verhindert, daß das neuronale Netz nur auf eine spezielle Abfolge von Eingabemustern trainiert wird.
Das Abbruchkriterium für das Training von neuronalen Netzen ist üblicherweise das Erreichen eines minimalen Fehlerwertes (Summe der quadratischen Abweichungen) zwischen den berechneten Ausgabewerten und den „korrekten“ Ausgabewerten. In dieser Arbeit wird hingegen eine Funktion verwendet, welche die Vorhersagegenauigkeit des neuronalen Netzes in Bezug auf die Erkennung von bekannten Bindungsregionen auf der Proteinoberfläche beurteilt: Diese Funktion kombiniert die Erkennungsraten S_α (Gleichung 6.2) der vier einzelnen Klassen von Teiloberflächen und die Gesamterkennungsrate S_{gesamt} zu einem Wert. Die Gesamterkennungsrate gibt den Anteil von Teiloberflächen an, die insgesamt durch die jeweilige Methode richtig als Bindungsstelle oder nichtbindende Teiloberfläche vorhergesagt werden.

$$S_{gesamt} = \frac{m_{Prot.} + m_{DNA} + m_{Lig.} + m_{n.bind.}}{n_{gesamt}} = \\ = \frac{n_{Prot.}}{n_{gesamt}} \cdot S_{Prot.} + \frac{n_{DNA}}{n_{gesamt}} \cdot S_{DNA} + \frac{n_{Lig.}}{n_{gesamt}} \cdot S_{Lig.} + \frac{n_{n.bind.}}{n_{gesamt}} \cdot S_{n.bind.} \tag{6.13}$$

mit

S_{gesamt}: Erkennungsrate aller Teiloberflächenarten zusammen

n_{gesamt}: Gesamtanzahl der untersuchten Teiloberflächen

Die kombinierte Erkennungsrate berechnet sich wie folgt:

$$S_{komb.} = \frac{1}{6} \cdot \left(S_{gesamt} + S_{Prot.} + S_{DNA} + S_{Lig.} + 2 \cdot S_{n.bind.} \right) \tag{6.14}$$

mit

$S_{komb.}$: Kombinierte Erkennungsrate des neuronalen Netzwerkes

Während des Trainings kann ein gegenläufiges Verhalten der Erkennungsraten der Bindungsbereiche und der nichtbindenden Oberflächen beobachtet werden. Hohe Erkennungsraten der Bindungsbereiche ($S_{Prot.}$, S_{DNA}, $S_{Lig.}$) sind mit einer niedrigeren Erkennungsrate der nichtbindenden Oberflächenbereiche ($S_{n.bind.}$) verbunden und umgekehrt. Dabei zeigt es sich, daß eine minimale Erhöhung der Bindungsbereicherkennungsraten oft mit einer nicht erwünschten, starken Erniedrigung von $S_{n.bind.}$ einhergeht. Um diesen Effekt in der

kombinierten Erkennungsrate $S_{komb.}$ zu berücksichtigen, wird die Erkennungsrate der nichtbindenden Oberflächen um den Faktor 2 stärker gewichtet. Das Training wird so lange durchgeführt, bis sich die kombinierte Erkennungsrate $S_{komb.}$ nicht mehr verbessert.
Die Eingabedaten werden in einen Trainings- und einen Testdatensatz aufgeteilt. Mit dem Trainingsdatensatz wird das Netz geschult, und der Testdatensatz dient zur Überprüfung der Vorhersagegenauigkeit. Üblicherweise werden die vorhandenen Daten gleichmäßig auf die beiden Datensätze verteilt. Erste Trainingsversuche mit einem Trainingssatz, der aus der Hälfte der 1,2 Millionen Oberflächensegmente bzw. Ein- und Ausgabemuster bestand, führten jedoch zu einem langsamen Lernverhalten und einer schlechten Vorhersagegenauigkeit des neuronalen Netzes. Deswegen wurde der Trainingssatz in mehreren Schritten verkleinert. Dazu wurde in jedem Schritt die Hälfte der Eingabemuster aus dem Trainingsdatensatz entfernt. Die Auswahl der zu entfernenden Eingabemuster geschah dabei zufällig. Mit abnehmender Anzahl von Teiloberflächen stieg sowohl die Geschwindigkeit des Lernvorganges als auch die Erkennungsleistung des Netzes an. Bei der Optimierung des Trainingsdatensatzes zeigte sich außerdem, daß die Anteile der drei Bindungsbereiche und der nichtbindenden Teiloberflächen im Trainingssatz ähnlich groß sein müssen. Sollte dies nicht der Fall sein, spezialisiert sich das neuronale Netz auf die Erkennung des Bindungsbereiches, der den größten Anteil im Trainingsdatensatz besitzt. Tabelle 5.13 zeigt, daß die Verteilung der verschiedenen Bindungsbereichtypen in den Teiloberflächen sehr unterschiedlich ist und daß der Hauptteil der Oberflächensegmente zu nichtbindenden Bereichen der Proteinoberfläche gehört. Um ein optimales Lernverhalten der neuronalen Netze zu gewährleisten, wird ein Trainingssatz von etwa 14.000 Oberflächensegmenten zusammengestellt, in dem der Anteil der Bindungsbereiche (insbesondere der DNA- und Ligandbindungsstellen) im Verhältnis zu den nichtbindenden Oberflächenbereichen stark erhöht ist. Zusätzlich enthält der Datensatz nur Oberflächensegmente, die gut in die vier Gruppen eingeordnet werden können. Bindende Oberflächensegmente müssen einen Flächenanteil von mindestens 40% an einem Bindungsbereich haben, während die nichtbindenden Segmente sogar zu mindestens 90% aus Oberflächenbereichen ohne Bindung zu anderen Molekülen bestehen müssen. Der Trainingsdatensatz basiert auf den Teiloberflächen des Enzym-Inhibitor-Datensatzes sowie einer zufälligen Auswahl von bindenden Teiloberflächen (DNA- und Ligandbindungsbereiche) aus dem Gesamtdatensatz. Insgesamt besteht der Trainingsdatensatz aus 13994 Teiloberflächen (57% nichtbindende Oberflächensegmente, 25% Protein-, 9% DNA- und 9% Ligandbindungsbereiche). Der Testdatensatz, der zur Überprüfung der Vorhersage-

genauigkeit nach jedem Lernschritt mittels Gleichung 6.14 verwendet wird, enthält die Oberflächensegmente des Gesamtdatensatzes, die nicht im Trainingssatz verwendet werden.

6.2.4 Trainingsergebnisse der neuronalen Netze

Nach jedem Trainingsschritt werden die Erkennungsraten der einzelnen Teiloberflächenklassen (Protein-, DNA-, Ligandbindungsbereich und nichtbindend) und die kombinierte Erkennungsrate (Gleichung 6.14) für den Trainings- und den Testdatensatz berechnet. In Diagramm 6.2 ist das Lernverhalten der beiden neuronalen Netze wiedergegeben (kombinierte Erkennungsrate im Testdatensatz). Die Lernkurve von beiden Netzwerken steigt schon nach wenigen hundert Lernschritten nicht mehr an. Bei beiden neuronalen Netzwerken wird deshalb das Training nach 500 Schritten abgebrochen. Netzwerk I erreicht schon nach wenigen Lernschritten (30 bis 40) seine maximale Leistungsfähigkeit. Der Lernprozeß des zweiten, etwas komplexeren Netzwerkes benötigt mehr Trainingsschritte. Erst nach etwa 150 bis 200 Lernschritten ist kein weiterer Aufwärtstrend in der Lernkurve zu erkennen. Dafür ist die durchschnittliche kombinierte Erkennungsrate von Netz II etwas höher als die von Netz I.

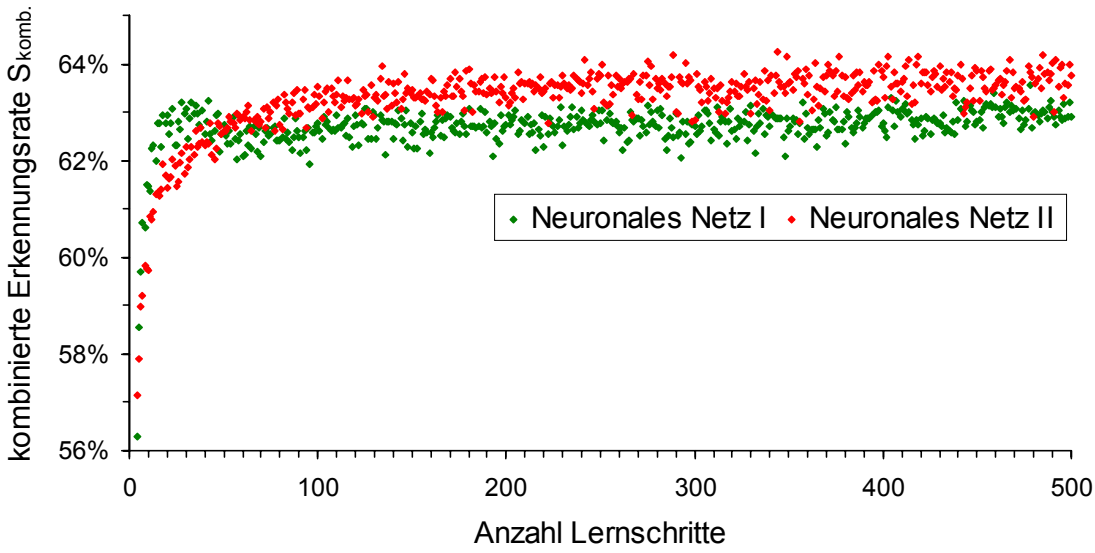


Diagramm 6.2: Lernverhalten der zwei neuronalen Netze. Für jeden Lernschritt wurde die kombinierte Erkennungsrate $S_{komb.}$ mit dem Testdatensatz berechnet.

Diagramm 6.3 zeigt den Lernfortschritt des neuronalen Netzes II anhand der Erkennungsraten der vier verschiedenen Teiloberflächenklassen (Protein-, DNA-, Ligandbindungsstellen und nichtbindende Oberflächen). Die entsprechenden Lernkurven von Netz I sind in Diagramm 9.2 im Anhang dargestellt.

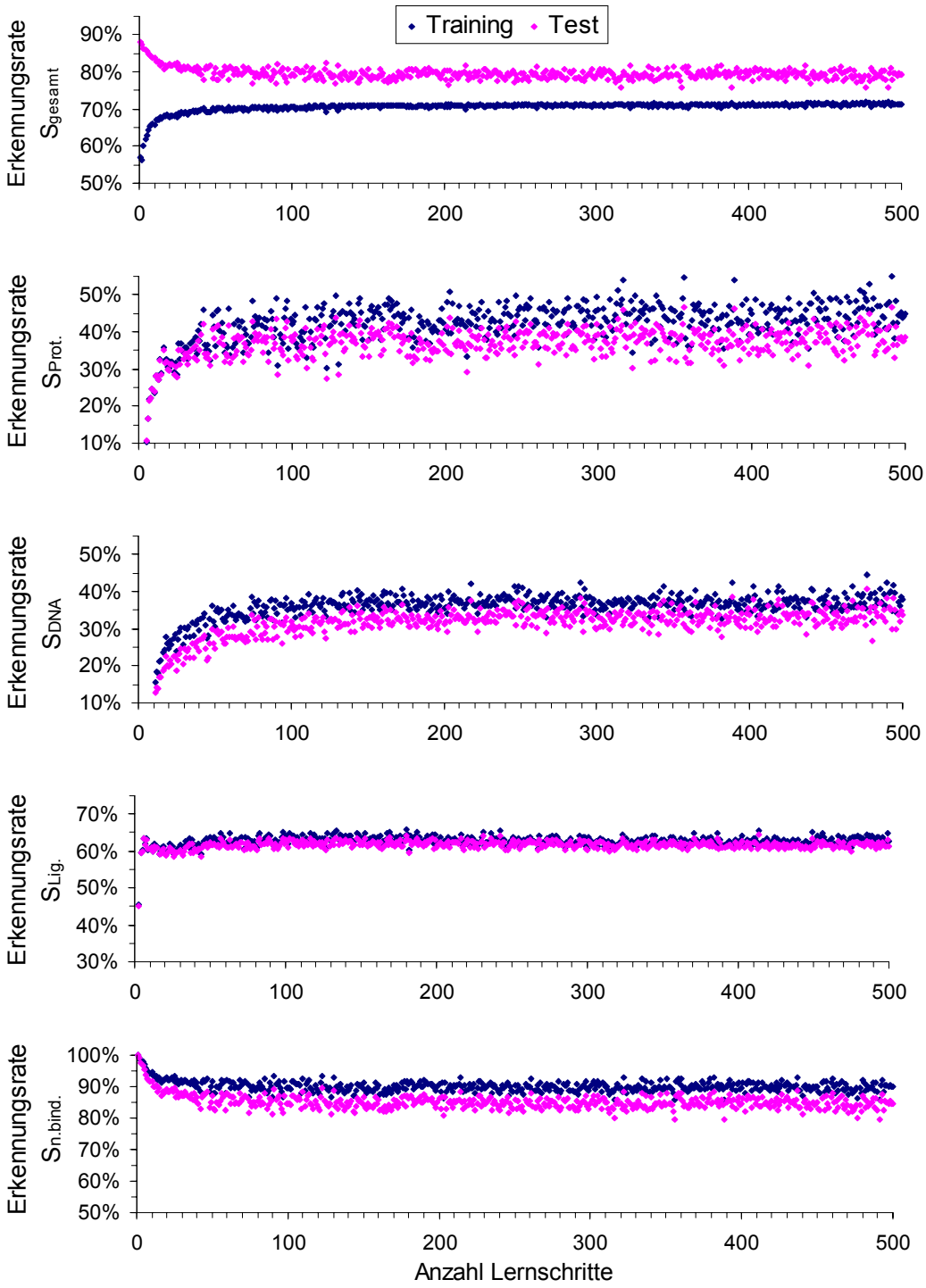


Diagramm 6.3: Lernverhalten des neuronalen Netzes II im Trainings- und Testdatensatz: Dargestellt sind von oben nach unten die Gesamterkennungsrate S_{gesamt} und die Erkennungsraten der vier Oberflächenklassen (Protein-, DNA-, Ligandbindungsbereich und nichtbindende Oberfläche).

Die Erkennungsraten der Protein-Protein- und Protein-DNA-Bindungsbereiche sind am Anfang des Lernprozesses sehr gering und steigen dann aber schnell an. Im Gegensatz dazu erkennt das neuronale Netz die nichtbindenden Oberflächen im wenig trainierten Zustand zu fast 100%. Mit Fortschreiten des Lernverfahrens fällt die Erkennungsrate $S_{n.bind.}$ dann auf ca. 90% ab. Dieses Verhalten ist auf die Zusammensetzung des Trainingsdatensatzes zurückzuführen. Er besteht zu fast 60% aus nichtbindenden Teiloberflächen. Zu Beginn des Trainings erkennt das neuronale Netz die Teiloberflächen, die den Hauptteil des Trainingssatzes ausmachen, am besten. Das ist auch der Grund für den fallenden Verlauf der Gesamterkennungsrate S_{gesamt} im Testdatensatz. Dort sind fast 90% der Teiloberflächen als nichtbindend klassifiziert. Am Anfang des Trainings werden diese Teiloberflächen sehr gut erkannt, und der hohe Anteil im Datensatz schlägt sich in der Gesamterkennungsrate S_{gesamt} nieder und bewirkt den abfallenden Kurvenverlauf. Dies ist ein zusätzlicher Grund, warum anstatt der Gesamt- die kombinierte Erkennungsrate $S_{komb.}$ als Abbruchkriterium des Lernverfahrens verwendet wird. Die Erkennungsraten verbessern sich während des Lernprozesses nicht kontinuierlich. Sie schwanken innerhalb eines gewissen Rahmens um einen Mittelwert. Dieser steigt langsam an und erreicht nach etwa 100-200 Lernschritten sein Maximum. Besonders starke Schwankungen der Erkennungsrate sind bei der Vorhersage der Protein-Protein- und Protein-DNA-Bindungsbereiche zu beobachten. Die Vorhersagegenauigkeit der Protein-Protein-Bindungsstellen im Testdatensatz pendelt zwischen 30% und 45%. Die Lernkurve der Protein-DNA-Bindungsregionen schwankt zwischen 25% und 35%. Die Erkennungsrate $S_{Lig.}$ der Ligandbindungsstellen schwankt nur sehr wenig (60% bis 65%), und der Maximalwert wird schon nach wenigen (etwa zehn) Schritten erreicht.

Das neuronale Netz I erreicht in Lernschritt 478 seine maximale kombinierte Erkennungsrate von 63,5%. Der Maximalwert von Netz II ist 64,2% und wird in Schritt 344 erreicht. In Tabelle 6.6 sind die Vorhersageleistungen der beiden neuronalen Netzwerke aufgeführt. Etwa ein Drittel der Protein- und DNA-Bindungsstellen, zwei Drittel der Ligandbindungsstellen und 85% der nichtbindenden Oberflächenbereiche im Testdatensatz werden von den neuronalen Netzen korrekt erkannt. Wie in den Lernkurven (Diagramm 6.3) sichtbar ist, werden in anderen Trainingsschritten teilweise höhere Erkennungsraten für einzelne Bindungsbereiche erreicht. Im Allgemeinen wird jedoch ein neuronales Netz benötigt, welches die verschiedenen Bindungsbereiche möglichst gleich gut vorhersagen kann. Das Training wird für beide Netze nach den genannten Lernschritten (478 bzw. 344) beendet.

Tabelle 6.6: Vorhersagegenauigkeit des neuronalen Netzes I und II nach Lernschritt 478 bzw. 344.

Erkennungsrate:	Neuronales Netz I		Neuronales Netz II	
	Trainings-datensatz	Test-datensatz	Trainings-datensatz	Test-datensatz
Protein-Protein-Bindungsstellen ($S_{Prot.}$)	43,6%	36,3%	43,7%	37,7%
Protein-DNA-Bindungsstellen (S_{DNA})	42,2%	36,8%	41,2%	37,4%
Protein-Ligand-Bindungsstellen ($S_{Lig.}$)	63,9%	62,1%	63,1%	62,4%
nichtbindende Oberflächen ($S_{n.bind.}$)	89,4%	83,8%	89,2%	84,4%
alle Oberflächenbereiche zusammen (S_{gesamt})	71,3%	78,3%	71,1%	79,0%
kombinierte Erkennungsrate ($S_{komb.}$)	66,6%	63,5%	66,2%	64,2%

Der Vergleich der Werte beider Netze zeigt, daß das neuronale Netz II etwas höhere Erkennungsraten besitzt. Die Unterschiede sind jedoch nur gering. Die Erkennungsraten des Test- und Trainingsdatensatzes unterscheiden sich ebenfalls nur wenig (maximal 7%). Daraus läßt sich schließen, daß die unterschiedlichen Arten von Teiloberflächen im Testdatensatz bzw. Gesamtdatensatz sehr gut durch die im Trainingsdatensatz enthaltenen Teiloberflächen repräsentiert werden, obwohl der Trainingsdatensatz erheblich kleiner ist (ca. 1/60 des Testdatensatzes). Wegen der etwas besseren Vorhersagegenauigkeit werden die weiteren Untersuchungen mit Netz II vorgenommen.

6.2.5 Eigenschaften der klassifizierten Teiloberflächen

Außer durch die Untersuchung der Erkennungsraten kann die korrekte Funktionsweise des neuronalen Netzes auch durch Analyse der molekularen Eigenschaften der mit dem Netz klassifizierten Teiloberflächen überprüft werden. Dazu werden mit dem neuronalen Netz II alle Teiloberflächen des Gesamtdatensatzes in die vier Oberflächenklassen eingeteilt und die Mittelwerte der molekularen Eigenschaften in den vier Klassen berechnet und mit den Ergebnissen aus Kapitel 5.7 (bzw. Tabellen 9.8 bis 9.11 im Anhang) verglichen. Wenn diese Werte übereinstimmen, ist gewährleistet, daß das neuronale Netz die auf die Teiloberflächen projizierten molekularen Eigenschaften mit den Bindungseigenschaften verknüpfen kann. Tabelle 6.7 enthält die angesprochenen Mittelwerte der von dem Netz klassifizierten Teiloberflächen. Die Werte stimmen mit den Ergebnissen von Kapitel 5.7 sehr gut überein und weichen nur geringfügig in einigen Eigenschaften voneinander ab.

Tabelle 6.7: Mittelwerte und Standardabweichungen der molekularen Eigenschaften der Teiloberflächen unterteilt nach der Klassifizierung durch das neuronale Netzwerk II.

	nichtbindend	Protein bindend	DNA bindend	Ligand bindend
Elektrostat. Pot. [kcal/(mol·e)]	-4,5 ± 32,8	-3,8 ± 27,6	75,4 ± 24,4	1,9 ± 36,9
Lokale Lipophilie	-0,072 ± 0,031	-0,021 ± 0,041	-0,067 ± 0,030	-0,014 ± 0,049
Flexibilität [$Å^2$]	26,3 ± 16,7	24,4 ± 17,5	28,2 ± 19,1	21,7 ± 14,0
H-Akz.-/H-Donordichte [$Å^{-2}$]	0,0454 ± 0,0097	0,0323 ± 0,0104	0,0491 ± 0,0099	0,0512 ± 0,0236
H-Akzeptordichte [$Å^{-2}$]	0,0239 ± 0,0054	0,0174 ± 0,0059	0,0146 ± 0,0054	0,0271 ± 0,0130
H-Donatordichte [$Å^{-2}$]	0,0216 ± 0,0078	0,0149 ± 0,0072	0,0344 ± 0,0071	0,0242 ± 0,0141
Tiefeninformation [Å]	1,05 ± 0,54	1,47 ± 1,02	1,74 ± 1,25	6,71 ± 2,47
Oberflächenkrümmung	2,30 ± 0,25	2,37 ± 0,37	2,27 ± 0,30	1,32 ± 0,44

6.2.6 Optimierung der Vorhersage von unbekannten Bindungsbereichen

Die Erkennung der verschiedenen Oberflächenklassen kann durch die Zusammensetzung des Trainingsdatensatzes gesteuert werden. Je höher der Anteil einer Oberflächenklasse am Trainingsdatensatz ist, desto mehr wird das neuronale Netz auf die Erkennung dieses Typs von Oberflächen trainiert und dessen Erkennungsrate steigt an. Das neuronale Netz kann also auf eine bestimmte Aufgabe hin optimiert werden. Wenn z.B. möglichst viele neue (noch nicht bekannte) Bindungsstellen gefunden werden sollen, muß der Anteil der nicht-bindenden Oberflächen im Trainingsdatensatz reduziert werden. Je mehr Bindungsbereiche vorhergesagt werden, desto höher ist jedoch auch die fehlerhafte Zuordnung von nicht-bindenden Bereichen zu einer der drei Bindungsklassen. Es gilt deshalb, eine Balance zwischen korrekter Erkennung von bekannten nichtbindenden Oberflächenbereichen und nicht bekannten Bindungsbereichen zu finden.

Der Gesamtdatensatz enthält auch viele einzelne Proteine, die mit anderen Molekülen Komplexe eingehen können. So sind unter den Einzelproteinen der *Protein Data Bank* auch sehr viele Strukturen, die in anderen Einträgen Komplexe mit weiteren Proteinen oder Molekülen bilden. Die gesamte molekulare Oberfläche dieser Proteine ist aber durch die Untersuchung der Teiloberflächen in Kapitel 5.7 als nichtbindend klassifiziert worden. Das neuronale Netz sollte trotzdem in der Lage sein, diese potentiellen Bindungsstellen vorherzusagen. Im nächsten Kapitel (6.3) werden Beispiele für solche erfolgreichen Vorhersagen präsentiert.

Das neuronale Netz II erkennt insgesamt 15% der als nichtbindend eingestuften Teil-oberflächen des Gesamtdatensatzes als mögliche Bindungsstellen. Von diesen ca. 150000

Teiloberflächen werden von dem neuronalen Netz 65% als Protein-, 18% als DNA und 17% als Ligandbindungsstellen klassifiziert.

Eine Verbesserung der Erkennung solcher unbekannten Bindungsstellen auf den Proteinoberflächen erfordert über die in dieser Arbeit vorgestellten Methoden hinausgehende Maßnahmen bei der Erstellung des Trainingsdatensatzes. So müßte sichergestellt werden, daß keine Teiloberflächen, welche in anderen Komplexen an der Bindung beteiligt sind, jedoch in der untersuchten Struktur als nichtbindend eingestuft werden, in den Trainingsdatensatz übernommen werden. Diese Bearbeitung des Datensatzes ist jedoch nicht mit automatisierten Mitteln durchführbar.

6.2.7 Einbindung des neuronalen Netzes in bestehende Programme

Mit dem SNNS-Programmpaket ist es möglich, ein trainiertes neuronales Netz in eine Programmroutine umzuwandeln. So kann die durch das neuronale Netz erreichte Vorhersage in ein eigenständiges Programm integriert werden, welches in das in Kapitel 4 vorgestellte Verfahren zur Untersuchung von Proteinen ausgehend von deren dreidimensionaler Struktur eingebaut wird. Dadurch wird es möglich, für neue Proteinstrukturen schnell und einfach eine Analyse der molekularen Oberfläche inklusive Vorhersage von möglichen Bindungsstellen durchzuführen. Da das neuronale Netz II eine etwas bessere Vorhersageleistung erzielt, wird nur dieses Netz für das Untersuchungsverfahren verwendet.

6.2.8 Darstellung der Bindungsbereichvorhersage auf der Proteinoberfläche

Neben der Vorhersage, ob einzelne Teiloberflächen Bindungen zu anderen Molekülen ausbilden können, ist es auch interessant, diese möglichen Bindungsbereiche direkt auf der molekularen Oberfläche anzuzeigen. Dazu muß die Ausgabeinformation (Vorhersage) des neuronalen Netzes auf die Punkte der Proteinoberfläche abgebildet werden. Die Abbildung der Bindungseigenschaft erfolgt für jede der verschiedenen Bindungsbereichklassen einzeln. Es ergeben sich somit drei neue Eigenschaften (Protein-, DNA- und Ligandbindungsmöglichkeit), die zur visuellen Untersuchung auf der molekularen Oberfläche dargestellt werden können. Die Berechnung dieser Bindungseigenschaft geschieht wie folgt: Für jeden Punkt der molekularen Oberfläche ist bekannt, zu welcher Teiloberfläche er gehört. So werden die mit dem neuronalen Netz ermittelten Bindungsanteile (Aktivierung der Ausgabeneuronen) allen Punkten der entsprechenden Teiloberfläche zugeordnet. Wenn

ein Punkt an mehreren Teiloberflächen beteiligt ist, werden die einzelnen Bindungsanteile aufsummiert und durch die Anzahl der Teiloberflächen, zu denen der Punkt gehört, geteilt:

$$BV_{x,i} = \frac{1}{n} \sum_{j=1}^{n} a_{x,j} \tag{6.15}$$

mit

$BV_{x,i}$: Bindungsbereichsvorhersage des Ausgabeneurons x für Oberflächenpunkt i

$a_{x,j}$: Aktivierung des Ausgabeneurons x für das Eingabemuster der Teiloberfläche j

n: Anzahl der Teiloberflächen, die Punkt i enthalten

6.3 Anwendungsbeispiele

In diesem Abschnitt wird die Leistungsfähigkeit des neuronalen Netzes (II) anhand von fünf repräsentativen Beispielen demonstriert. Dazu werden aus dem Gesamtdatensatz vier Proteinkomplexe und ein Einzelprotein ausgewählt und mit dem in Kapitel 4 beschriebenen Verfahren untersucht. Anschließend werden mit Hilfe des neuronalen Netzes mögliche Bindungsbereiche der Proteine bestimmt. Da durch die dreidimensionale Struktur der kristallisierten Komplexe die korrekten Bindungsbereiche der untersuchten Proteine bekannt sind, kann die Vorhersagegenauigkeit des Netzes mit dieser Information überprüft werden. Dazu wird die Bindungsbereichsvorhersage auf die molekulare Oberfläche der Proteine, wie in Kapitel 6.2.8 beschrieben, abgebildet und visuell mit der kristallisierten Komplexstruktur verglichen. Es werden folgenden Komplexe vorgestellt:

- Protein-DNA-Komplex: Tumorsuppressor p53 / DNA (1tup) [140]
- Protein-Ligand-Komplex: Dihydrofolat-Reduktase / Methotrexat (4dfr) [141]
- Protein-Protein-Komplex: Thymidilat-Kinase-Dimer (2tmk) [142]
- Protein-Protein-Komplex: β-Trypsin / BPTI (2ptc) [143]
- Protein: BPTI (4pti) [143]

Bei der Auswahl dieser Komplexe aus dem Gesamtdatensatz wurde darauf geachtet, daß keiner dieser Proteinkomplexe im Trainingsdatensatz des neuronalen Netzes enthalten ist. Damit soll nochmals gezeigt werden, daß das neuronale Netz in der Lage ist, auch mögliche Bindungsbereiche von neuen, unbekannten Proteinen zu erkennen.

6.3.1 Protein-DNA-Komplex des Tumorsuppressorproteins p53

Das Protein p53 ist ein wichtiges Protein im Zellteilungszyklus. Wird die DNA durch Umwelteinflüsse beschädigt, steigt die p53-Konzentration in der Zelle an. Der Zellzyklus wird dann in der G1/S-Phase gestoppt, und die körpereigenen Reparaturmechanismen können die DNA reparieren, oder der programmierte Zelltod (Apoptose) wird ausgelöst, um die Vermehrung geschädigter Zellen zu verhindern Diese biologische Funktion ist eng mit der Fähigkeit, an spezifische DNA-Sequenzen zu binden, verknüpft [140,144-148].
In Abbildung 6.4 ist die Kristallstruktur des Protein-p53-DNA-Komplexes dargestellt. Das Protein bindet in der großen Furche der DNA (Stabmodell im Vordergrund). Auf der molekularen Oberfläche des Proteins ist der vom neuronalen Netz vorhergesagte DNA-Bindungsbereich farbig dargestellt (rot = DNA-Bindungsbereich). Die Vorhersage gibt die Position der beobachteten DNA-Bindungsstelle der Komplexkristallstruktur exakt wieder.

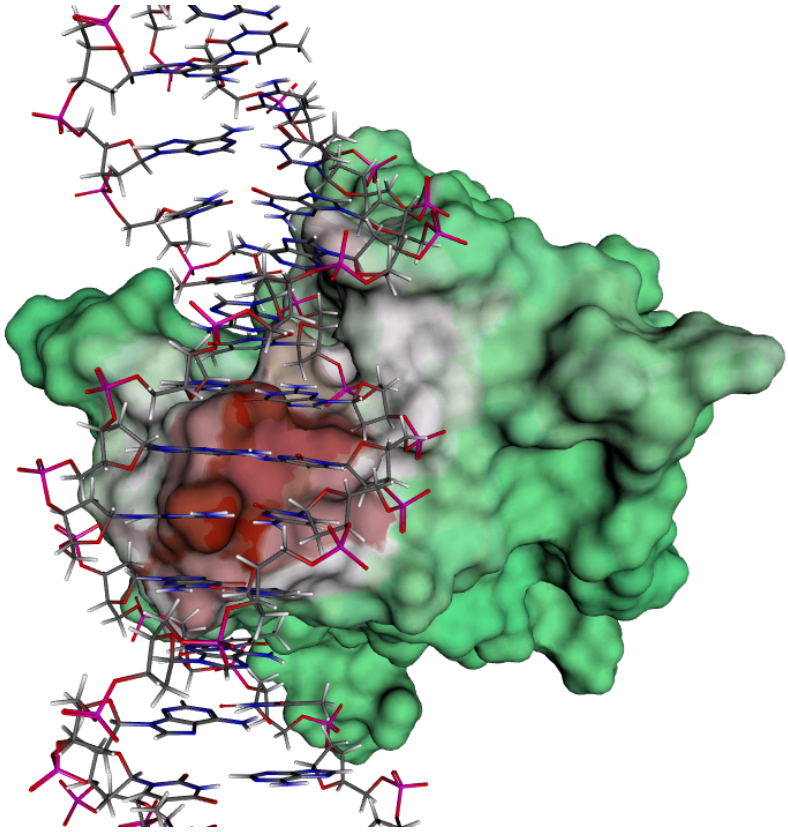


Abbildung 6.4: Vorhersage des DNA-Bindungsbereiches des Proteins p53 mit Hilfe des neuronalen Netzes II. Die DNA-Bindungsbereichvorhersage (Gleichung 6.15) ist auf der Proteinoberfläche des p53-Proteins farbig dargestellt (rot: hohe Wahrscheinlichkeit für DNA-Bindungsstelle, grün: niedrige Wahrscheinlichkeit). Die Positionen von DNA und Protein entsprechen der Kristallstruktur.

6.3.2 Dihydrofolat-Reduktase komplexiert mit dem Inhibitor Methotrexat

Das Enzym Dihydrofolat-Reduktase ist an der Synthese der Nucleotidbase Thymin beteiligt. Eine der Vorstufen in der Synthese ist das Desoxythymidilat. Zu dessen Synthese wird der Cofaktor Tetrahydrofolat in Dihydrofolat umgesetzt, welches durch die Dihydrofolat-Reduktase wieder zu Tetrahydrofolat regeneriert wird. Der Inhibitor Methotrexat bindet im aktiven Zentrum der Dihydrofolat-Reduktase und blockiert damit die Regenerierung des Tetrahydrofolat. Dadurch ist auch die Synthese von Thymin behindert, was zu einer toxischen Wirkung führt [141].

Das aktive Zentrum der Reduktase befindet sich in einer tiefen Tasche der molekularen Oberfläche. In Abbildung 6.5 ist die Kristallstruktur des Enzyms mit Inhibitor dargestellt. Der Inhibitor Methotrexat ist als Kugel-Stab-Modell gezeigt und ist direkt in der Tasche positioniert. Auf der molekularen Oberfläche des Enzyms ist die Vorhersage des neuronalen Netzes, wie oben, dargestellt. Die gesamte Tasche wird von dem neuronalen Netz II als mögliche Ligandbindungsstelle erkannt und rot markiert. Wiederum stimmt der vorhergesagte Bindungsbereich mit der beobachteten Position des Liganden sehr gut überein.

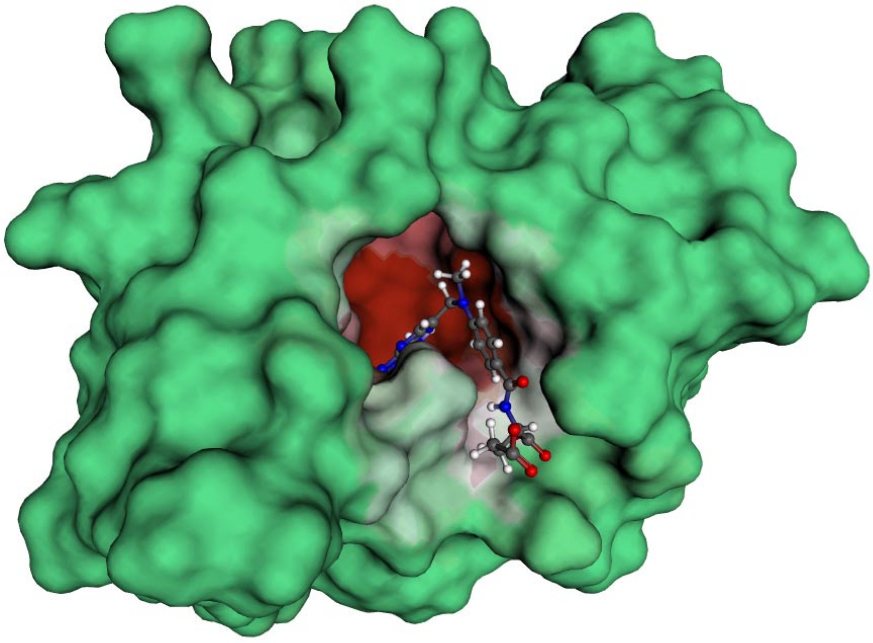


Abbildung 6.5: Vorhersage des Ligandbindungsbereiches auf der molekularen Oberfläche der Dihydrofolat-Reduktase mittels des neuronalen Netzes II. Die Vorhersage (Gleichung 6.15) ist auf der Proteinoberfläche des Enzyms farbig dargestellt (rot: hohe Wahrscheinlichkeit für Ligandbindungsstelle, grün: niedrige Wahrscheinlichkeit). Der Ligand Methotrexat (Kugel-Stab-Modell) bindet im kristallisierten Komplex (4dfr) genau im hier für die Reduktase vorhergesagten Bindungsbereich.

6.3.3 Thymidilat-Kinase-Dimer

Die Thymidilat-Kinase ist ein wichtiges Enzym für die Zellreproduktion. Sie phosphoryliert Desoxythymidinmonophospat (dTMP) unter Anwesenheit von Adenosintriphosphat und Magnesium zum entsprechenden Diphosphat (dTDP). Das Diphosphat wird anschließend durch die Nucleosiddiphosphat-Kinase in Desoxythymidintriphosphat (dTTP) umgesetzt, welches für die DNA-Synthese benötigt wird [142].

Die Thymidilat-Kinase bildet ein Dimer. Es ist aus zwei Proteinen mit identischer Aminosäuresequenz und Sekundärstruktur aufgebaut. Abbildung 6.6 zeigt die Kristallstruktur des Dimerkomplexes. Auf der molekularen Oberfläche des einen Proteins ist die Bindungsbereichvorhersage des neuronalen Netzes farbig dargestellt (rot = Protein-Protein-Bindungsstelle). Das neuronale Netz erkennt mehrere mögliche Protein-Protein-Bindungsregionen auf der molekularen Oberfläche. Der größte zusammenhängende Bereich befindet sich genau in der Protein-Protein-Bindungsregion des Dimers. Ein kleiner Bereich (in Abbildung 6.6 gelb markiert) wird dabei vom Netz nicht als Protein-Protein-Bindungsbereich identifiziert, sondern als nichtbindende Oberfläche klassifiziert. Der Hauptteil der Bindungsregion wird jedoch korrekt vorhergesagt.

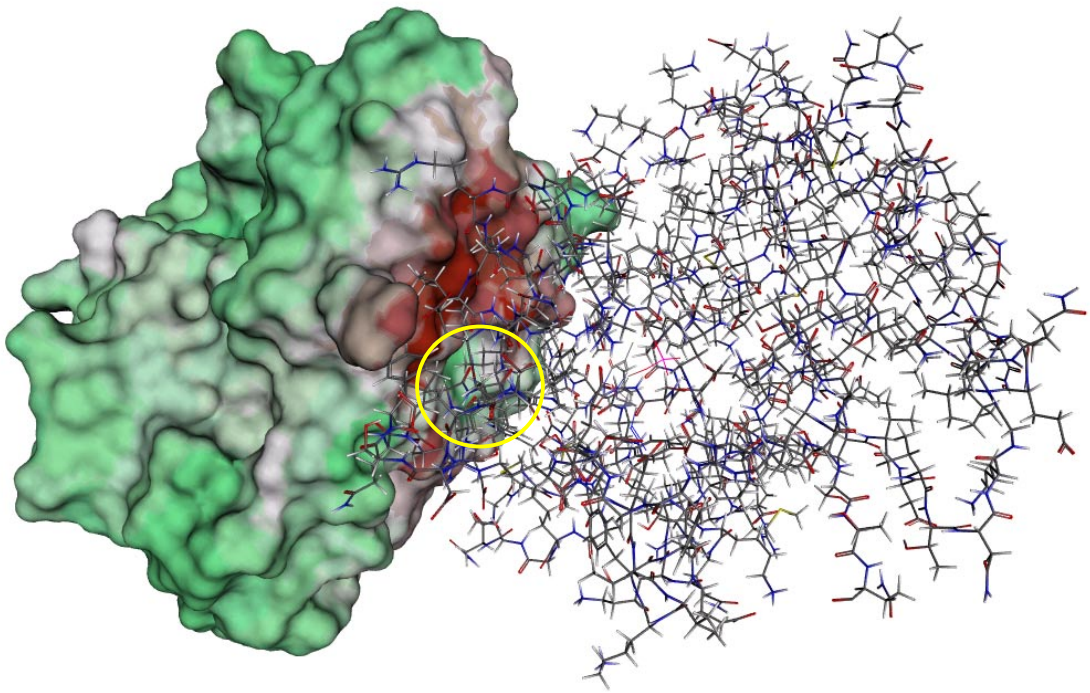


Abbildung 6.6: Vorhersage des Bindungsbereiches zwischen den beiden Proteinen des Thymidilat-Kinase-Dimers. Die Bindungsbereichvorhersage ist auf der molekularen Oberfläche des ersten Proteins farbig abgebildet (rot: hohe Wahrscheinlichkeit für Protein-Protein-Bindungsstelle, grün: niedrige Wahrscheinlichkeit). Das zweite Protein ist als Stabmodell dargestellt.

6.3.4 Enzym-Inhibitor-Komplex von β-Trypsin mit dem Protein PTI

Das Verdauungsenzym Trypsin gehört zur Klasse der Serinproteasen und dient zum Proteinabbau durch spezifische Spaltung von Peptidbindungen. Die Spezifität der bei der Peptidspaltung bevorzugten Seitenketten wird durch die Struktur der Spezifitätstasche in der Nähe des aktiven Zentrums gewährleistet. Trypsin katalysiert selektiv die Spaltung von Peptidbindungen an der Carboxylseite von Aminosäuren mit langgestreckten basischen Seitenketten (Lysin und Arginin). Um eine Verdauung körpereigener Eiweißmoleküle bei Substratmangel zu verhindern, wird die Aktivität des Trypsins durch den kompetitiven Inhibitor PTI (*Pancreatic Trypsin Inhibitor*) reguliert. PTI lagert sich so an die Trypsinoberfläche an, daß das aktive Zentrum für andere Moleküle blockiert wird [143].

Abbildung 6.7 zeigt, wie das Protein PTI (Stabmodell) in der Kristallstruktur des Enzym-Inhibitor-Komplexes genau in der Spezifitätstasche des Trypsins bindet. Die molekulare Oberfläche des Trypsins ist wiederum entsprechend der Bindungsbereichsvorhersage eingefärbt (rot = Protein-Protein-Bindungsstelle). Die Spezifitätstasche und die unmittelbare Umgebung wird von dem neuronalen Netz korrekt als mögliche Bindungsstelle für Proteine identifiziert. Genau in diesem Bereich lagert sich der Inhibitor an.

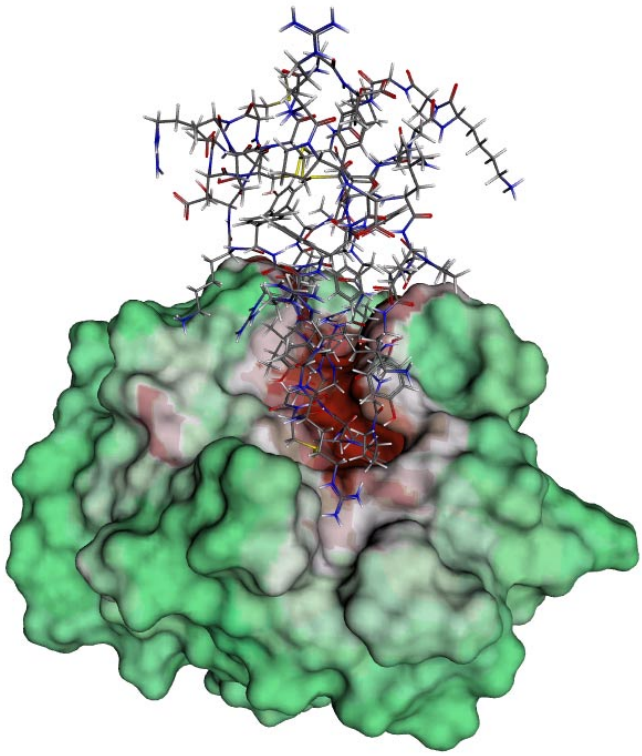


Abbildung 6.7: Vorhersage des Protein-Protein-Bindungsbereiches des Trypsins. Die Vorhersage des Bindungsbereiches ist auf der Oberfläche farbig dargestellt (rot: hohe Wahrscheinlichkeit für Protein-Protein-Bindungsstelle, grün: niedrige Wahrscheinlichkeit). Der Inhibitor PTI (als Stabmodel dargestellt) bindet genau in dieser Bindungsstelle.

6.3.5 BPTI (*Bovine Pancreatic Trypsin Inhibitor*)

Die meisten Proteine liegen ungebunden in einer anderen Konformation vor, als wenn sie mit anderen Molekülen einen Komplex bilden. Besonders die Seitenketten der Aminosäuren auf der molekularen Oberfläche sind beweglich und nehmen in Lösung unterschiedlichste Positionen ein. Durch die Komplexbildung werden Teile dieser Seitenketten fixiert. In manchen Fällen erfordert die Komplexbildung sogar eine erhebliche Änderung großer Teile der Proteinstruktur bzw. Proteinoberfläche. Dieser Vorgang wird als *Induced Fit* bezeichnet. Wenn die Komplexbildung mit einer solch starken Änderung einhergeht, ist eine Vorhersage der möglichen Bindungsbereiche auf der molekularen Oberfläche sehr schwierig. Wenn die Strukturänderungen jedoch nur geringfügig sind und die Form und Aminosäurenzusammensetzung der Oberfläche sich nur leicht ändern, sollte das neuronale Netz in der Lage sein, die Positionen von möglichen Bindungsbereichen zu bestimmen.

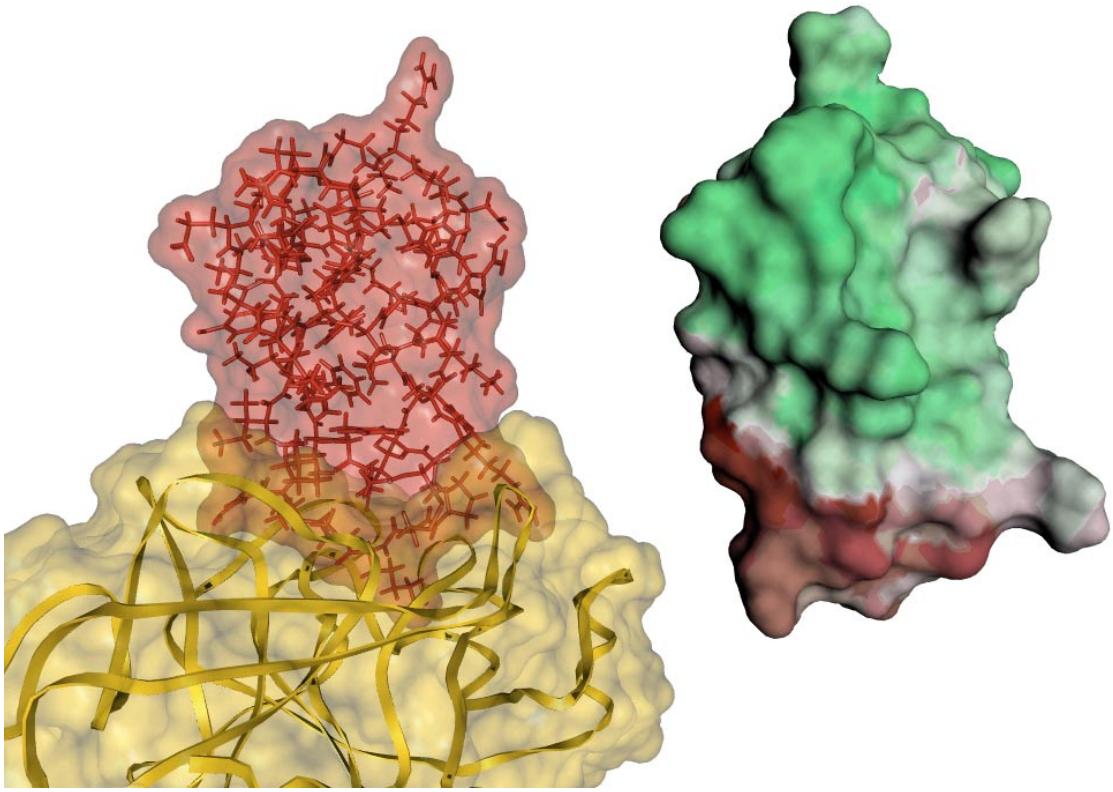


Abb. 6.8: Vorhersage des Protein-Protein-Bindungsbereiches des BPTI:
Links) Trypsin-BPTI-Komplex (2ptc): Sowohl Trypsin (gelb) als auch BPTI (rot) sind mit durchsichtigen Proteinoberflächen abgebildet, um die Orientierung der Moleküle im Komplex wiederzugeben.
Rechts) BPTI (4pti): Die Vorhersage des Protein-Protein-Bindungsbereiches durch das neuronale Netz II ist auf der Proteinoberfläche farbig dargestellt (rot: hohe Wahrscheinlichkeit für Protein-Protein-Bindungsstelle, grün: niedrige Wahrscheinlichkeit). Das Protein ist in der gleichen Orientierung wie das BPTI im Trypsin-Komplex dargestellt.

Als Testbeispiel dient hierzu das schon oben beschriebene Protein BPTI (*Bovine Pancreatic Trypsin Inhibitor*). In der *Protein Data Bank* gibt es neben mehreren Komplexen mit BPTI (z.B. der Trypsin-PTI-Komplex 2ptc) auch Einträge mit BPTI ohne Bindung zu anderen Proteinen (z.B. 4pti). In Abbildung 6.8 sind zwei BPTI-Varianten nebeneinander abgebildet (links: BPTI im Trypsin-BPTI-Komplex 2ptc; rechts: einzelnes BPTI aus Eintrag 4pti). Beide Proteine sind bezüglich ihrer Aminosäurekette aufeinander ausgerichtet und mit der molekularen Oberfläche dargestellt. Die Proteinoberfläche der beiden Moleküle zeigt deutliche Unterschiede. Auf der linken Seite der Abbildung ist erkennbar, wie das BPTI an das Trypsin bindet. Die Bindungsbereichsvorhersage für das einzelne Protein (4pti) ist auf der molekularen Oberfläche abgebildet (rot = Protein-Protein-Bindungsbereich). Der untere Teil des Proteins wird als Bindungsstelle identifiziert. Dieser Teil bindet im Trypsin-Komplex in der Spezifitätstasche des Enzyms. Das neuronale Netz ist also trotz der Unterschiede der Oberflächen der beiden BPTI-Strukturen in der Lage, nur auf der Basis der physikochemischen Eigenschaften den Oberflächenbereich zu identifizieren, der mit Trypsin eine Komplexbindung eingehen kann.

7 Zusammenfassung und Ausblick

In der hier vorliegenden Arbeit wird über die Entwicklung eines Verfahrens zur computergestützten Analyse der dreidimensionalen Strukturen von Proteinen und Proteinkomplexen berichtet. Mit diesem Verfahren können Proteinstrukturen komplett automatisch bearbeitet und analysiert werden. Dadurch war es im Gegensatz zu den bisherigen Untersuchungen [19,20,22,24-26,28-33,37] möglich, auch eine sehr große Anzahl (mehrere tausend) von Proteinstrukturen zu bearbeiten und eine umfassende Datenbasis für die nachfolgenden Untersuchungen zu schaffen. Mit der vorgestellten Methode wurden die molekularen Eigenschaften und Oberflächen der Proteinmoleküle ausgehend von deren dreidimensionalen Strukturen berechnet. Anschließend wurden die Eigenschaften zur weiteren Auswertung auf die molekularen Oberflächen projiziert. Für die Berechnungen wurde die komplette Strukturinformation (Positionen aller Atome) der einzelnen Proteine benötigt. Da diese Informationen für viele Proteine nicht vollständig bekannt sind, enthält das Verfahren eine Methode zur Ergänzung fehlender Atompositionen. Die weiteren Untersuchungen konzentrierten sich auf die Proteinoberflächen und die Bindungsbereiche der Proteinkomplexe. Dabei wurde zwischen Protein-, DNA- und Ligandbindungsbereichen unterschieden. In dieser Arbeit wurden alle Struktureinträge der *Brookhaven Protein Data Bank* (10213 Strukturen - Stand Oktober 1999) mit dem vorgestellten Verfahren analysiert. Insgesamt konnten für 7821 (82%) Proteine und Proteinkomplexe in der Datenbank die molekularen Eigenschaften und Oberflächen berechnet und untersucht werden.

Bei der Analyse der Daten wurde der Gesamtdatensatz der 7821 fehlerfrei bearbeiteten Strukturen in drei weitere Datensätze unterteilt, um zu untersuchen, ob bestimmte Arten von Proteinen bzw. Proteinkomplexen spezifische physikochemische Eigenschaften oder Oberflächencharakteristika besitzen, welche die Wechselwirkungen der Proteine mit anderen Molekülen beeinflussen. Zusätzlich zu einem Datensatz aus Antigen-Antikörper-Komplexen und einem aus Enzym-Inhibitor-Komplexen wurde der reduzierte Datensatz von Hobohm et al. [119] verwendet. Die vergleichenden Untersuchungen zeigten, daß der reduzierte Datensatz trotz der geringeren Anzahl von Strukturen (ca. 400) die Eigenschaften des Gesamtdatensatzes sehr gut repräsentiert.

Aus den Untersuchungen der Bindungsbereiche der Proteinkomplexe folgt, daß es keine Beziehung zwischen der Proteingröße bzw. -masse und der Größe der Bindungsbereiche an der molekularen Oberfläche gibt. Die Größe ist nur von dem jeweiligen Komplex und der Funktion der Bindung abhängig. Die durchschnittliche Größe der Protein-Protein-

Bindungsregionen beträgt ca. 780 Å^2. Die DNA-Bindungsbereiche sind deutlich kleiner (ca. 480 Å^2), die Größe der Ligandbindungsstellen ist im Mittel nur 130 Å^2 groß.

Die Aminosäurenzusammensetzung der Proteine ist nicht gleichmäßig. An der Proteinoberfläche ist der Anteil an hydrophilen Aminosäuren höher als über das gesamte Protein betrachtet. Die hydrophoben Aminosäuren sind dagegen eher im Proteininneren anzutreffen. Die unterschiedlichen Bindungsbereiche der Proteine (Protein, DNA und Ligand) unterscheiden sich ebenfalls in ihrer Zusammensetzung. Die Bindungsbereiche zu anderen Proteinen ähneln mehr dem Proteininneren als der Außenseite, d.h. es sind deutlich mehr hydrophobe Aminosäuren vorhanden. Die hydrophoben Aminosäuren nehmen dort auch einen deutlich höheren Anteil an der Oberfläche ein als in den anderen Bereichen der molekularen Proteinoberfläche. In den DNA-Bindungsstellen sind dagegen besonders viel Arginin und Lysin vorhanden. Diese beiden Aminosäuren sind positiv geladen und können Salzbrücken zu den negativ geladenen Phosphatgruppen von DNA-Molekülen ausbilden, wodurch eine starke Komplexbindung möglich wird. Negativ geladene Aminosäuren (Glutaminsäure und Asparaginsäure), die diese Wechselwirkung stören würden, sind in den DNA-Bindungsstellen der Proteinoberfläche selten.

Die unterschiedlichen Zusammensetzungen (Aminosäurenverteilung) in den Bindungsregionen und nichtbindenden Bereichen der Proteinaußenseiten sind auch für Unterschiede der lokalen chemischen Eigenschaften in diesen Oberflächenbereichen verantwortlich. An der Proteinaußenseite sind viele mögliche Wasserstoffakzeptoren und -donatoren vorhanden. Etwa jedes vierte bis fünfte Atom an der Oberfläche ist ein Wasserstoffdonator oder -akzeptor, das entspricht etwa 2,3 Wasserstoffakzeptoren bzw. -donatoren pro 100 Å^2 Oberfläche. Der Anteil an Akzeptoren ist geringfügig höher als der Anteil der Donatoren. In den Protein-Protein-Bindungsregionen ist die Dichte der Akzeptoren und Donatoren geringer als in den nichtbindenden Bereichen. Die DNA-Bindungsbereiche zeichnen sich durch einen sehr hohen Anteil von Wasserstoffakzeptoren aus. Bei der Untersuchung der Abstände zwischen Wasserstoffakzeptoren und -donatoren an der molekularen Oberfläche wurde deutlich, daß sich Verteilungsmuster an der Oberfläche bilden. Es gibt bevorzugte Abstände zwischen den Akzeptoren und Donatoren. Diese sind hauptsächlich durch die Anordnung der Atome in funktionellen Gruppen wie z.B. Amin-, Carboxyl- oder Ammoniumgruppen innerhalb einer einzelnen Aminosäure bedingt. Jedoch bilden sich auch Muster (bevorzugte Abstände) zwischen den Akzeptoren bzw. Donatoren verschiedener Aminosäuren aus.

In den Protein-Protein-Komplexen wurden die Kontaktwechselwirkungen zwischen den Proteinen untersucht (Aminosäurenkontakte zwischen den Komplexpartnern). Es zeigte sich, daß es kein allgemeingültiges Muster von Aminosäurenkontakten gibt. Die Kontaktwechselwirkungen sind stark von der Funktion der einzelnen Komplexbindung abhängig, jedoch lassen sich aus dem Vergleich der Kontaktmatrizen mit den numerisch berechneten Kontaktverhältnissen einige allgemeine Trends ableiten: So treten hydrophobe Aminosäuren wie z.B. Leucin häufiger in Kontakt miteinander, als aufgrund ihres Anteils an der Proteinoberfläche im Bindungsbereich zu erwarten ist. Salzbrücken zwischen den positiv und negativ geladenen Aminosäuren (Arginin und Lysin bzw. Glutaminsäure und Asparaginsäure) sind ebenfalls sehr oft zu finden. In den Bindungsbereichen sind auch viele Kontakte bzw. Disulfidbrücken zwischen Cysteinen der beiden Komplexpartner vorhanden. Wechselwirkungen zwischen hydrophoben und hydrophilen Aminosäuren in den Bindungsbereichen sind hingegen seltener anzutreffen.

Für jedes der bearbeiteten Proteine wurden sechs verschiedene charakteristische Eigenschaften berechnet und auf deren molekulare Oberflächen projiziert: Elektrostatisches Potential, lokale Lipophilie, Wasserstoffdonatoren- bzw. Wasserstoffakzeptorendichte, molekulare Flexibilität, Tiefeninformation und Oberflächenkrümmung. Darauf aufbauend wurden die Proteinoberflächen mit einem auf *Fuzzy Logic* basierenden Verfahren in sich überlappende Teiloberflächen eingeteilt und die Mittelwerte der molekularen Eigenschaften dieser Teiloberflächen ausgewertet. Jede dieser Teiloberflächen ist etwa 250 $Å^2$ groß, das entspricht etwa der Fläche von zwei bis drei Aminosäuren an der Proteinoberfläche. Aus der Analyse der sechs Eigenschaften in den drei verschiedenen Bindungsbereichen und den nichtbindenden Oberflächenbereichen konnten interessante Ergebnisse gewonnen werden: Die Eigenschaften der Protein-Protein-Bindungsstellen ähneln den Werten der nichtbindenden Bereiche der Proteinoberfläche. Jedoch unterscheidet sich die lokale Lipophilie der beiden Oberflächenbereiche deutlich. Die Protein-Protein-Bindungsstellen sind im Mittel, wie bereits erwähnt, deutlich hydrophober als die nichtbindenden Bereiche und ähneln somit mehr dem Inneren der Proteine als deren Außenseite. Die DNA-Bindungsbereiche zeichnen sich besonders durch ihre überwiegend elektropositive Polarisierung und die hohe Dichte an Wasserstoffdonatoren aus. Dadurch ist eine starke Wechselwirkung mit DNA-Molekülen möglich. Die Ligandbindungsstellen befinden sich zu einem großen Teil in tiefen Taschen und Höhlen der Proteinoberfläche. Die Oberfläche ist dort meist konkav gekrümmt und hydrophober als die nichtbindende Oberfläche an der Außenseite der Proteine.

Ausgehend von diesen Untersuchungsergebnissen wurden zwei Methoden zur Vorhersage von Bindungsbereichen an den Proteinoberflächen entwickelt und getestet. Die Verfahren basieren dabei nur auf den physikalischen und chemischen Eigenschaften, die den Teiloberflächen zugeordnet wurden. Die erste Methode versucht, die Bindungseigenschaft der Teiloberflächen mittels einer einfachen Zielfunktion (Linearkombination der Eigenschaftswerte) abzubilden. Nach der Parametrisierung zeigte sich jedoch, daß mit dieser Vorgehensweise nur die Ligandbindungsbereiche von Proteinen zufriedenstellend bestimmt werden können. Protein- und DNA-Bindungsstellen konnten mit der vorgestellten Zielfunktion nicht vorhergesagt werden. Deshalb wurde eine zweite Vorhersagemethode entwickelt, die auf einem neuronalen Netz basiert. Das neuronale Netz wurde mit einem stark reduzierten Satz von Teiloberflächen trainiert und konnte die durch die obigen Untersuchungen bekannten Bindungsstellen aller untersuchten Proteinkomplexe zufriedenstellend identifizieren und als Protein-, DNA- oder Ligandbindungsbereiche klassifizieren. An einigen Beispielen wurde gezeigt, daß das neuronale Netz auch in der Lage ist, mögliche Bindungsstellen von nicht im Trainingsdatensatz enthaltenen Proteinen vorherzusagen.

Das in der vorliegenden Arbeit vorgestellte Verfahren zur Untersuchung von biologisch relevanten Makromolekülen ist bisher auf die Analyse von Proteinstrukturen beschränkt. Durch die Verbesserung der bestehenden Programme oder Einbeziehung weiterer Programme zur Ergänzung unvollständiger Strukturinformationen kann das Verfahren auf die Untersuchung anderer nichtproteinogener Molekülstrukturen ausgeweitet werden. Dadurch wird es auch möglich, die Strukturen der Proteindatenbank zu analysieren, die das hier vorgestellte Verfahren nicht bearbeiten konnte. Solche Verbesserungen ermöglichen auch die differenziertere Behandlung der verschiedenen Bindungsbereiche, so daß z.B. die Ligandbindungsstellen nach Molekültyp des gebundenen Liganden getrennt untersucht werden können. Weiterhin sollten die präsentierten Untersuchungsmethoden in zukünftigen Arbeiten auf die seit Ende 1999 neu in die Proteindatenbank aufgenommenen Strukturen (über 7000 neue Strukturen) angewendet werden.

Die sechs untersuchten Eigenschaften sind nur ein kleiner Teil der physikochemischen Eigenschaften, die auf die molekularen Oberflächen projiziert und ausgewertet werden können. In weiterführenden Untersuchungen können zusätzliche molekulare Eigenschaften berechnet und ausgewertet werden. Dadurch werden weitere Informationen über die Wechselwirkungen von Proteinen und Proteinkomplexen gewonnen. Auch bei der Analyse

der Bindungsbereiche gibt es noch eine Reihe offenstehende Fragen, die durch anknüpfende Untersuchungen beantwortet werden sollten: Welche Form haben die Bindungsbereiche? Werden Bindungsbereiche durch bestimmte Abfolgen von Aminosäuren gebildet? Unterscheidet sich die Packungsdichte der Atome im Bindungsbereich und im Proteininneren? Diese Ideen erfordern die Entwicklung neuer Berechnungs- und Auswertungsmethoden. Hier könnten auch neuronale Netze wie z.B. von Zupan und Gasteiger vorgeschlagen [10] sehr hilfreich sein. Die neuen Methoden können dann sehr einfach in das hier beschriebene Verfahren integriert und auf die gesamte Proteindatenbank angewendet werden.

Sehr interessant ist sicherlich auch eine Unterscheidung und Bewertung der Protein-Protein-Bindungsbereiche bezüglich ihres Typs und der Stärke der Komplexbindungen. Da Komplexbindungskonstanten nur von wenigen Strukturen in der Proteindatenbank bekannt sind, müssen hierzu Verfahren zur Abschätzung der Konstanten eingesetzt werden [66].

Die vorgestellte Methode zur Vorhersage von Bindungsbereichen basiert auf einem einfachen neuronalen Netz. Durch die Wahl von komplexeren neuronalen Netzen und einer umfassenden Optimierung des Netzaufbaus bzw. der Lernparameter kann die Erkennungsleistung sicherlich gesteigert werden. Andere Untersuchungen lassen den Schluß zu, daß durch ein solches Vorgehen starke Verbesserungen der Vorhersagequalität von neuronalen Netzen erreicht werden können [10,14-16]. Die Veränderungen sollten dabei alle Aspekte von neuronalen Netzwerken abdecken: Anzahl der Neuronen, Verknüpfungen zwischen den Neuronen, Lernregel, Lernparameter, etc. Auch die Wahl eines anderen Netzwerktyps, wie z.B. des in der Arbeitsgruppe von Gasteiger verwendeten Kohonen-Netzwerkes [12], kann zur Verbesserung der Vorhersageleistung beitragen. Die Wahl der Eingabedaten hat ebenfalls entscheidenden Einfluß auf die Leistung der Vorhersagemethode. Die Optimierungsmöglichkeiten erstrecken sich also auch auf die Wahl der durchschnittlichen Teiloberflächengröße, die auf die Oberflächen projizierten Eigenschaften, die Parameter der Oberflächensegmentierung und die Wahl der Eigenschaften, die zur Segmentierung herangezogen werden. Die Bindungsbereichsvorhersage stützt sich in dieser Arbeit nur auf die eine, gerade untersuchte Teiloberfläche und die darauf projizierten Eigenschaften. Jedoch hat sicherlich die direkte Umgebung dieses Teilbereiches Einfluß auf die Bindungsfähigkeit. In zukünftigen Vorhersagemethoden sollten deshalb, vergleichbar mit der von Exner vorgestellten Strategie zur Vorhersage von Bindungsmodi von Proteinkomplexen [64,116,117], die direkt angrenzenden Teiloberflächen mit berücksichtigt werden.

Das präsentierte neuronale Netz zur Vorhersage von Bindungsbereichen ist nur eine von vielen Anwendungsmöglichkeiten des hier vorgestellten automatisierten Verfahrens. Die Untersuchungsergebnisse können z.B. auch für die Verbesserung oder Neuentwicklung von Methoden in anderen Gebieten wie z.B. dem Protein-Docking, rationalem Wirkstoff-design oder Proteindesign verwendet werden. All diese Ideen zeigen, daß mit diesem Verfahren eine Tür zu einer großen Anzahl von Möglichkeiten der Analyse von Proteinen bzw. Proteinkomplexen und ihren Eigenschaften aufgestoßen wurde.

8 Literatur

1. D. Voet und J.G. Voet; Biochemie; Verlag Chemie; Weinheim, **1992**.

2. E. Fischer; Einfluss Der Configuration Auf Die Wirkung Der Enzyme; *Chem.Ber.* (**1894**) *27*; 2985-2993.

3. H.-J. Böhm, G. Klebe und H. Kubinyi; Wirkstoffdesign; Spektrum Akademischer Verlag GmbH; Heidelberg, **1996**.

4. F.C. Bernstein, T.F. Koetzle, G.J. Williams, E.E. Meyer, Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi und M. Tasumi; The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures; *J.Mol.Biol.* (**1977**) *112*(3); 535-542.

5. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov und P.E. Bourne; The Protein Data Bank; *Nucleic Acids Res.* (**2000**) *28*(1); 235-242.

6. S.K. Burley, S.C. Almo, J.B. Bonanno, M. Capel, M.R. Chance, T. Gaasterland, D. Lin, A. Sali, F.W. Studier und S. Swaminathan; Structural Genomics: Beyond the Human Genome Project; *Nat.Genet.* (**1999**) *23*(2); 151-157.

7. W. Heiden; Methoden zur computerunterstützten Untersuchung selektiver Oberflächeneigenschaften von Proteinen; *Dissertation*; Darmstadt (**1993**).

8. B. Lee und F.M. Richards; The Interpretation of Protein Structures: Estimation of Static Accessibility; *J.Mol.Biol.* (**1971**) *55*; 379-400.

9. F.M. Richards; Areas, Volumes, Packing, and Protein Structure; *Annu.Rev.Biophys. Bio.* (**1977**) *6*; 151-176.

10. J. Zupan und J. Gasteiger; Neural Networks for Chemists: An Introduction; VCH; Weinheim, **1993**.

11. G. Schneider und P. Wrede; Artificial Neural Networks for Computer-Based Molecular Design; *Prog.Biophys.Mol.Biol.* (**1998**) *70*; 175-222.

12. S. Anzali, G. Barnickel, M. Krug, J. Sadowski, M. Wagener, J. Gasteiger und J. Polanski; The Comparison of Geometric and Electronic Properties of Molecular Surfaces by Neural Networks: Application to the Analysis of Corticosteroid-Binding Globulin Activity of Steroids; *J.Comput.Aided Mol.Des* (**1996**) *10*(6); 521-534.

13. J. Sadowski und H. Kubinyi; A Scoring Scheme for Discriminating Between Drugs and Nondrugs; *J.Med.Chem.* (**1998**) *41*(18); 3325-3329.

14. N. Qian und T.J. Sejnowski; Predicting the Secondary Structure of Globular Proteins Using Neural Network Models; *J.Mol.Biol.* (**1988**) *202*(4); 865-884.

15. B. Rost und C. Sander; Prediction of Protein Secondary Structure at Better Than 70% Accuracy; *J.Mol.Biol.* (**1993**) *232*(2); 584-599.

16. M. Reczko; Protein Secondary Structure Prediction With Partially Recurrent Neural Networks; *SAR QSAR.Environ.Res.* (**1993**) *1*(2-3); 153-159.

17. A.H. Elcock, D. Sept und J.A. McCammon; Computer Simulation of Protein-Protein Interactions; *J.Phys.Chem.B* (**2001**) *105*; 1504-1518.

18. C. Chothia und J. Janin; Principles of Protein-Protein Recognition; *Nature* (**1975**) *256*(5520); 705-708.

19. J. Janin, S. Miller und C. Chothia; Surface, Subunit Interfaces and Interior of Oligomeric Proteins; *J.Mol.Biol.* (**1988**) *204*(1); 155-164.

20. J. Janin und C. Chothia; The Structure of Protein-Protein Recognition Sites; *J.Biol.Chem.* (**1990**) *265*(27); 16027-16030.

21. S. Jones und J.M. Thornton; Protein-Protein Interactions: a Review of Protein Dimer Structures; *Prog.Biophys.Mol.Biol.* (**1995**) *63*(1); 31-65.

22. S. Jones und J.M. Thornton; Principles of Protein-Protein Interactions; *P.Natl.Acad. Sci.USA* (**1996**) *93*(1); 13-20.

23. S.L. Lin und R. Nussinov; Molecular Recognition Via Face Center Representation of a Molecular Surface; *J.Mol.Graphics* (**1996**) *14*(2); 78-7.

24. S. Jones und J.M. Thornton; Analysis of Protein-Protein Interaction Sites Using Surface Patches; *J.Mol.Biol.* (**1997**) *272*(1); 121-132.

25. S. Jones und J.M. Thornton; Prediction of Protein-Protein Interaction Sites Using Patch Analysis; *J.Mol.Biol.* (**1997**) *272*(1); 133-143.

26. C.J. Tsai, S.L. Lin, H.J. Wolfson und R. Nussinov; Studies of Protein-Protein Interfaces: a Statistical Analysis of the Hydrophobic Effect; *Protein Sci.* (**1997**) *6*(1); 53-64.

27. A.A. Bogan und K.S. Thorn; Anatomy of Hot Spots in Protein Interfaces; *J.Mol.Biol.* (**1998**) *280*(1); 1-9.

28. L.L. Conte, C. Chothia und J. Janin; The Atomic Structure of Protein-Protein Recognition Sites; *J.Mol.Biol.* (**1999**) *285*(5); 2177-2198.

29. F. Glaser, D.M. Steinberg, I.A. Vakser und N. Ben Tal; Residue Frequencies and Pairing Preferences at Protein-Protein Interfaces; *Proteins* (**2001**) *43*(2); 89-102.

30. A.J. McCoy, E. Chandana, V und P.M. Colman; Electrostatic Complementarity at Protein/Protein Interfaces; *J.Mol.Biol.* (**1997**) *268*(2); 570-584.

31. D. Xu, C.J. Tsai und R. Nussinov; Hydrogen Bonds and Salt Bridges Across Protein-Protein Interfaces; *Protein Eng* (**1997**) *10*(9); 999-1012.

32. D. Xu, S.L. Lin und R. Nussinov; Protein Binding Versus Protein Folding: the Role of Hydrophilic Bridges in Protein Associations; *J.Mol.Biol.* (**1997**) *265*(1); 68-84.

33. L. Young, R.L. Jernigan und D.G. Covell; A Role for Surface Hydrophobicity in Protein-Protein Recognition; *Protein Sci.* (**1994**) *3*; 717-729.

34. M. Scarsi, N. Majeux und A. Caflisch; Hydrophobicity at the Surface of Proteins; *Proteins* (**1999**) *37*(4); 565-575.

35. L.F. Pacios; Distinct Molecular Surfaces and Hydrophobicity of Amino Acid Residues in Proteins; *J.Chem.Inf.Comp.Sci.* (**2001**) *41*(5); 1427-1435.

36. K.A. Dill; Dominant Forces in Protein Folding; *Biochemistry* (**1990**) *29*(31); 7133-7155.

37. S. Jones, P. van Heyningen, H.M. Berman und J.M. Thornton; Protein-DNA Interactions: A Structural Analysis; *J.Mol.Biol.* (**1999**) *287*(5); 877-896.

38. R. Wang, L. Liu, L. Lai und Y. Tang; SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex; *J.Mol.Model.* (**1998**) *4*; 379-394.

39. G.P. Brady und P.F.W. Stouten; Fast Prediction and Visualization of Protein Binding Pockets With PASS; *J.Comput.Aid.Mol.Des.* (**2000**) *14*; 383-401.

40. D.J. Danziger und P.M. Dean; Automated Site-Directed Drug Design: a General Algorithm for Knowledge Acquisition About Hydrogen-Bonding Regions at Protein Surfaces; *P.Roy.Soc.Lond.B Bio.* (**1989**) *236*(1283); 101-113.

41. K.P. Peters, J. Fauck und C. Frömmel; The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using Only Geometric Criteria; *J.Mol.Biol.* (**1996**) *256*(1); 201-213.

42. M.L. Verdonk, J.C. Cole, P. Watson, V. Gillet und P. Willett; SuperStar: Improved Knowledge-Based Interaction Fields for Protein Binding Sites; *J.Mol.Biol.* (**2001**) *307*(3); 841-859.

43. R.A. Laskowski, N.M. Luscombe, M.B. Swindells und J.M. Thornton; Protein Clefts in Molecular Recognition and Function; *Protein Sci.* (**1996**) *5*(12); 2438-2452.

44. J. Ruppert, W. Welch und A.N. Jain; Automatic Identification and Representation of Protein Binding Sites for Molecular Docking; *Protein Sci.* (**1997**) *6*(3); 524-533.

45. C.M. Oshiro, I.D. Kuntz und M.A. Knegtel; Molecular Docking and Structure-Based Design; in *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons; Chichester, **1998**; Band 3, 1606-1613.

46. I. Muegge und M. Rarey; Small Molecule Docking and Scoring; in *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH; New York, **2001**; Band 17, 1-60.

47. L.P. Ehrlich und R.C. Wade; Protein-Protein-Docking; in *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH; New York, **2001**; Band 17, 61-97.

48. D.G. Levitt und L.J. Banaszak; POCKET: a Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids; *J.Mol.Graphics* (**1992**) *10*(4); 229-234.

49. J.S. Delaney; Finding and Filling Protein Cavities Using Cellular Logic Operations; *J.Mol.Graphics* (**1992**) *10*(3); 174-7, 163.

50. R.A. Laskowski; SURFNET: A Program for Visualizing Molecular Surfaces, Cavities, and Intermolecular Interactions; *J.Mol.Graphics* (**1995**) *13*; 323-330.

51. E.C. Meng, B.K. Shoichet und I.D. Kuntz; Automated Docking With Grid-Based Energy Evaluation; *J.Comp.Chem.* (**1992**) *13*(4); 505-525.

52. T. Exner, M. Keil, G. Moeckel und J. Brickmann; Identification of Substrate Channels and Protein Cavities; *J.Mol.Model.* (**1998**) *4*; 340-343.

53. H. Edelsbrunner und E.P. Mücke; Three Dimensional Alpha-Shapes; *ACM Transact. Graph.* (**1994**) *13*(1); 43-72.

54. P.J. Goodford; A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules; *J.Med.Chem.* (**1985**) *28*; 849-857.

55. G.E. Kellogg, S.F. Semus und D.J. Abraham; HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA; *J.Comput.Aid.Mol.Des.* (**1991**) *5*; 545-552.

56. G. Klebe; The Use of Composite Crystal-Field Environments in Molecular Recognition and the De Novo Design of Protein Ligands; *J.Mol.Biol.* (**1994**) *237*(2); 212-235.

57. R.A. Laskowski, J.M. Thornton, C. Humblet und J. Singh; X-SITE: Use of Empirically Derived Atomic Packing Preferences to Identify Favourable Interaction Regions in the Binding Sites of Proteins; *J.Mol.Biol.* (**1996**) *259*(1); 175-201.

58. M. Hendlich, F. Rippmann und G. Barnickel; LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins; *J.Mol.Graph. Model.* (**1997**) *15*(6); 359-363.

59. R.M. MacCallum, A.C. Martin und J.M. Thornton; Antibody-Antigen Interactions: Contact Analysis and Binding Site Topography; *J.Mol.Biol.* (**1996**) *262*(5); 732-745.

60. O. Lichtarge, H.R. Bourne und F.E. Cohen; An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families; *J.Mol.Biol.* (**1996**) *257*(2); 342-358.

61. A. Armon, D. Graur und N. Ben Tal; ConSurf: an Algorithmic Tool for the Identification of Functional Regions in Proteins by Surface Mapping of Phylogenetic Information; *J.Mol.Biol.* (**2001**) *307*(1); 447-463.

62. F. Pazos, M. Helmer-Citterich, G. Ausiello und A. Valencia; Correlated Mutations Contain Information About Protein-Protein Interaction; *J.Mol.Biol.* (**1997**) *271*(4); 511-523.

63. P. Aloy, E. Querol, F.X. Aviles und M.J. Sternberg; Automated Structure-Based Prediction of Functional Sites in Proteins: Applications to Assessing the Validity of Inheriting Protein Function From Homology in Genome Annotation and to Protein Docking; *J.Mol.Biol.* (**2001**) *311*(2); 395-408.

64. T. Exner; Computergestützte Strukturbestimmung biochemischer Komplexe durch einen Fuzzy Logic-basierten Algorithmus; *Dissertation*; Darmstadt (**2000**).

65. M. Keil; Computergestützte Untersuchungen der Bindungsregion des p53-DNA-Komplexes; *Diplomarbeit*; Darmstadt (**1996**).

66. T. Borosch; Computergestützte Analyse der Wechselwirkung zwischen Cystatin und einem künstlichen Antikörper; *Diplomarbeit*; Darmstadt (**1999**).

67. J. Brickmann, T. Exner, M. Keil, R. Marhöfer und G. Moeckel; Molecular Models: Visualization; in *The Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer III, H. F., Schreiner, P. R., Eds.; John Wiley & Sons; Chichester, **1998**; Band 3, 1679-1693.

68. P.G. Mezey; Molecular Surfaces; in *Reviews in Computational Chemistry*; **1991**; 265-294.

69. R. Corey und L. Pauling; Molecular Models of Amino Acids, Peptides and Proteins; *Rev.Sci.Instr.* (**1953**) *24*; 612.

70. M.L. Connolly; Solvent-Accessible Surfaces of Proteins and Nucleic Acids; *Science* (**1983**) *221*; 709-713.

71. M.L. Connolly; Analytical Molecular Surface Calculation; *J.Appl.Crystallogr.* (**1983**) *16*; 548-558.

72. M.M. Francl, R.F. Hout Jr und W.J. Hehre; Representation of Electron Densities. 1. Sphere Fits to Total Electron Density Surfaces; *J.Am.Chem.Soc.* (**1984**) *106*; 563-570.

73. J.-P. Doucet und J. Weber; Computer-Aided Molecular Design: Theory and Applications; Academic Press; London San Diego New York Boston Sydney Tokyo, **2001**.

74. W. Heiden, T. Goetze und J. Brickmann; Fast Generation of Molecular Surfaces From 3D Data Fields With an Enhanced 'Marching Cube' Algorithm; *J.Comp.Chem.* (**1993**) *14*; 246-250.

75. R.R. Gabdoulline und R.C. Wade; Analytically Defined Surfaces to Analyze Molecular Interaction Properties; *J.Mol.Graphics* (**1996**) *14*(6); 341-345.

76. A. Bondi; Van Der Waals Volumes and Radii; *J.Phys.Chem.* (**1964**) *68*(3); 441-451.

77. W. Heiden, M. Schlenkrich und J. Brickmann; Triangulation Algorithms for the Representation of Molecular Surface Properties; *J.Comput.Aid.Mol.Des.* (**1990**) *4*(3); 255-269.

78. A.H. Juffer und H.J. Vogel; A Flexible Triangulation Method to Describe the Solvent-Accessible Surface of Biopolymers; *J.Comput.Aid.Mol.Des.* (**1998**) *12*(3); 289-299.

79. W. Cai, M. Zhang und B. Maigret; New Approach for Representation of Molecular Surface; *J.Comp.Chem.* (**1998**) *19*(16); 1805-1811.

80. S.L. Chan und E.O. Purisima; Molecular Surface Generation Using Marching Tetrahedra; *J.Comp.Chem.* (**1998**) *19*(11); 1268-1277.

81. M.L. Connolly; The Molecular Surface Package; *J.Mol.Graphics* (**1993**) *11*(2); 139-141.

82. M.F. Sanner und A.J. Olson; Reduced Surface: An Efficient Way to Compute Molecular Surfaces; *Biopolymers* (**1996**) *38*; 305-320.

83. M. Totrov und R. Abagyan; The Contour-Buildup Algorithm to Calculate the Analytical Molecular Surface; *J.Struct.Biol.* (**1996**) *116*(1); 138-143.

84. R. Zauhar; SMART: A Solvent-Accessible Triangulated Surface Generator for Molecular Graphics and Boundary Element Applications; *J.Comput.Aid.Mol.Des.* (**1995**) *9*; 149-159.

85. W. Lorensen und H. Cline; A High Resolution 3D Surface Construction Algorithm; *Computer Graphics* (**1987**) *21*; 163-169.

86. J. Brickmann, T. Goetze, W. Heiden, G. Moeckel, S. Reiling, H. Vollhardt und C.D. Zachmann; Interactive Visualization of Molecular Scenarios With MOLCAD/SYBYL; in *Data Visualisation in Molecular Science: Tools for Insight and Innovation*; Bowie, J. E., Ed.; Addison-Wesley Publishing Company Inc.; Reading, Mass., **1995**; 83-97.

87. J. Brickmann, W. Heiden, H. Vollhardt und C.D. Zachmann; New Man-Machine Communication Strategies in Molecular Modelling; in *Proceedings of the 28-Th Annual Hawaii International Conference on System Sciences. Biotechnology Computing*; Hunter, L., Shriver, B. D., Eds.; IEEE Computer Society Press; Los Alamitos, CA., **1995**; Band V, 273-282.

88. R.S. Bohacek und C. McMartin; Definition and Display of Steric, Hydrophobic, and Hydrogen-Bonding Properties of Ligand Binding Sites in Proteins Using Lee and Richards Accessible Surface: Validation of a High-Resolution Graphical Tool for Drug Design; *J.Med.Chem.* (**1992**) *35*(10); 1671-1684.

89. J. Warwicker und H.C. Watson; Calculation of the Electric Potential in the Active Site Cleft Due to Alpha-Helix Dipoles; *J.Mol.Biol.* (**1982**) *157* ; 671-679.

90. I. Klapper, R. Hagstrom, R. Fine, K.A. Sharp und B.H. Honig; Focusing of Electric Fields in the Active Site of Cu-Zn Superoxide Dismutase: Effects of Ionic Strength and Amino-Acid Modification; *Proteins* (**1986**) *1*; 47-59.

91. K.A. Sharp und B.H. Honig; Electrostatic Interactions in Macromolecules: Theory and Applications; *Annu.Rev.Biophys.Biophys.Chem.* (**1990**) *19*(301); 332.

92. M.K. Gilson, K.A. Sharp und B.H. Honig; Calculating the Electrostatic Potential of Molecules in Solution: Method and Error Assessment; *J.Comp.Chem.* (**1987**) *9*(4); 327-335.

93. R.E. Bruccoleri, J. Novotny, M.E. Davis und K.A. Sharp; Finite Difference Poisson-Boltzmann Electrostatic Calculations: Increased Accuracy Achieved by Harmonic Dielectric Smoothing and Charge Antialiasing; *J.Comp.Chem.* (**1997**) *18*(2); 268-276.

94. A. Nicholls und B.H. Honig; A Rapid Finite Difference Algorithm, Utilizing Successive Over-Relaxation to Solve the Poisson-Boltzmann Equation; *J.Comp. Chem.* (**1991**) *12*(4); 435-445.

95. B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, D.J. States, S. Swaminathan und M. Karplus; CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations; *J.Comp.Chem.* (**1983**) *4*; 187-217.

96. T. Fujita, J. Iwasa und C. Hansch; A New Substituent Constant, Π, Derived From Partition Coefficients; *J.Am.Chem.Soc.* (**1964**) *86*; 5175-5180.

97. H. Van de Waterbeemd und R. Mannhold; Lipophilicity Descriptors for Structure-Property Correlation Studies: Overview of Experimental and Theoretical Methods and a Benchmark of Log P Calculations; in *Lipophilicity in Drug Action and Toxicology, Volume 4*; Pliska, V., Testa, B., Eds.; VCH; Weinheim, **1996**; 401-418.

98. A.K. Ghose und G.M. Crippen; Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationship I. Partition Coefficients As a Measure of Hydrophobicity; *J.Comp.Chem.* (**1986**) *7*; 565-577.

99. A.K. Ghose, A. Pritchett und G.M. Crippen; Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationship III: Modeling Hydrophobic Interactions; *J.Comp.Chem.* (**1988**) *9*; 80-90.

100. V.N. Viswanadhan, A.K. Ghose, G.R. Revankar und R.K. Robins; Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationship. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occuring Nucleoside Antibiotics; *J.Chem.Inf.Comp.Sci.* (**1989**) *29*; 163-172.

101. E. Audry, J.P. Dubost, J.C. Colleter und P. Dallet; Une Nouvelle Approche Des Relations Structure-Activité: Le "Potentiel De Lipophilie Moleculaire"; *Eur.J.Med. Chem.* (**1986**) *21*; 71-72.

102. P. Furet, A. Sele und N.C. Cohen; 3D Molecular Lipophilicity Potential Profiles: a New Tool in Molecular Modeling; *J.Mol.Graphics* (**1988**) *6*; 182-189.

103. J.-L. Fauchère, P. Quarendon und L. Kaetterer; Estimating and Representing Hydrophobicity Potential; *J.Mol.Graphics* (**1988**) *6*; 203-206.

104. I. Rozas und M. Martín; Molecular Lipophilic Potential on Van Der Waals Surfaces As a Tool in the Study of 4-Alkylpyrazoles; *J.Chem.Inf.Comp.Sci.* (**1996**) *36*; 872-878.

105. A.J. Leo; Calculating Log P_{oct} From Structures; *Chem.Rev.* (**1993**) *93*; 1281-1306.

106. W. Heiden, G. Moeckel und J. Brickmann; A New Approach to Analysis and Display of Local Lipophilicity/Hydrophilicity Mapped on Molecular Surfaces; *J.Comput.Aid. Mol.Des.* (**1993**) *7*(5); 503-514.

107. M.P. Allen und D.J. Tildesley; Computer Simulation of Liquids; Claredon Press; Oxford, **1987**.

108. E.H. Spanier; Algebraic Topology; McGraw-Hill; New York, **1966**.

109. C.D. Zachmann, W. Heiden, M. Schlenkrich und J. Brickmann; Topological Analysis of Complex Molecular Surfaces; *J.Comp.Chem.* (**1992**) *13*(1); 76-84.

110. C.D. Zachmann; Methoden zur Quantifizierung der Rauheit und Flexibilität molekularer Oberflächen; *Dissertation*; Darmstadt (**1995**).

111. C.D. Zachmann, S.M. Kast und J. Brickmann; Quantification and Visualization of Molecular Surface Flexibility; *J.Mol.Graphics* (**1995**) *13*(2); 89-97.

112. F.H. Allen und O. Kennard; 3D Search and Research Using the Cambridge Structural Database; *Chemical Design Automation News* (**1993**) *8*(1); 31-37.

113. The RCSB Protein Data Bank; **http//www.rcsb.org/pdb**.

114. R. Jäger; Empirisches Berechnungsverfahren von Oberflächenbezogenen 1-Oktanol/Wasser-Verteilungskoeffizienten; Diplomarbeit; Darmstadt, **1996**.

115. W. Heiden und J. Brickmann; Segmentation of Protein Surfaces Using Fuzzy Logic; *J.Mol.Graphics* (**1994**) *12*(2); 106-115.

116. T. Exner, M. Keil und J. Brickmann; Pattern Recognition Strategies for Molecular Surfaces: I. Pattern Generation Using Fuzzy Set Theory; *J.Comp.Chem.* (**2002**) zur Veröffentlichung angenommen.

117. T. Exner, M. Keil und J. Brickmann; Pattern Recognition Strategies for Molecular Surfaces: II. Surface Complementarity; *J.Comp.Chem.* (**2002**) zur Veröffentlichung angenommen.

118. C.J. Tsai, S.L. Lin, H.J. Wolfson und R. Nussinov; A Dataset of Protein-Protein Interfaces Generated With a Sequence-Order-Independent Comparison Technique; *J.Mol.Biol.* (**1996**) *260*(4); 604-620.

119. U. Hobohm und C. Sander; Enlarged Representative Set of Protein Structures; *Protein Sci.* (**1994**) *3*; 522.

120. U. Hobohm, M. Scharf, R. Schneider und C. Sander; Selection of Representative Protein Data Sets; *Protein Sci.* (**1992**) *1*(3); 409-417.

121. N.M. Luscombe, R.A. Laskowski und J.M. Thornton; Amino Acid-Base Interactions: a Three-Dimensional Analysis of Protein-DNA Interactions at an Atomic Level; *Nucleic Acids Res.* (**2001**) *29*(13); 2860-2874.

122. D.R. Davies und G.H. Cohen; Interactions of Protein Antigens With Antibodies; *P.Natl.Acad.Sci.USA* (**1996**) *93*(1); 7-12.

123. E.J. Sundberg, M. Urrutia, B.C. Braden, J. Isern, D. Tsuchiya, B.A. Fields, E.L. Malchiodi, J. Tormo, F.P. Schwarz und R.A. Mariuzza; Estimation of the Hydrophobic Effect in an Antigen-Antibody Protein-Protein Interface; *Biochemistry* (**2000**) *39*(50); 15375-15387.

124. B. Pirard, G. Baudoux und F. Durant; A Database Study of Intermolecular NH···O Hydrogen Bonds for Carboxylates, Sulfonates and Monohydrogen Phosphonates; *Acta Crystallogr.B* (**1995**) *51*; 103-107.

125. SYBYL 6.5; Tripos Inc.; 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.

126. G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee und M.H. Zehfus; Hydrophobicity of Amino Acid Residues in Globular Proteins; *Science* (**1985**) *229*(4716); 834-838.

127. D.S. Goodsell und A.J. Olson; Soluble Proteins: Size, Shape and Function; *Trends Biochem.Sci.* (**1993**) *18*(3); 65-68.

128. F.K. Pettit und J.U. Bowie; Protein Surface Roughness and Small Molecular Binding Sites; *J.Mol.Biol.* (**1999**) *285*(4); 1377-1382.

129. F. Torrens, J. Sánchez-Marín und I. Nebot-Gil; New Dimension Indices for the Characterization of the Solvent-Accessible Surface; *J.Comp.Chem.* (**2001**) *22*(5); 477-487.

130. C.D. Zachmann, S.M. Kast, A. Sariban und J. Brickmann; Self-Similarity of Solvent-Accessible Surface of Biological and Synthetical Macromolecules; *J.Comp.Chem.* (**1993**) *14*(11); 1290-1300.

131. P. McCaldon und P. Argos; Oligopeptide Biases in Protein Sequences and Their Use in Predicting Protein Coding Regions in Nucleotide Sequences; *Proteins* (**1988**) *4*(2); 99-122.

132. B. Pullman; Electrostatics of Polymorphic DNA; *J.Biomol.Struct.Dyn.* (**1983**) *1*; 773-794.

133. H.-J. Böhm; The Development of a Simple Empirical Scoring Function to Estimate the Binding Constant for a Protein-Ligand Complex of Known Three-Dimensional Structure; *J.Comput.Aided Mol.Des* (**1994**) *8*(3); 243-256.

134. A. Ajay und M.A. Murcko; Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes; *J.Med.Chem.* (**1995**) *38*(26); 4953-4967.

135. A. Zell, N. Mache, T. Sommer und T. Korb; Design of the SNNS Neural Network Simulator; in *Östereichische Artificial-Intelligence-Tagung*; Springer Verlag; Wien, **1991**; 93-102.

136. A. Zell; Simulation Neuronale Netze; R. Oldenbourg Verlag; München Wien, **1997**.

137. R. Brause; Neuronale Netze: Eine Einführung in die Neuroinformatik; B.G. Teubner; Stuttgart, **1995**.

138. D.O. Hebb; The Organization of Behavior; John Wiley & Sons; New York, **1949**.

139. D.E. Rumelhart und J.L. McClelland; Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations; MIT Press; **1986**.

140. Y. Cho, S. Gorina, P.D. Jeffrey und N.P. Pavletich; Crystal Structure of P53 Tumor Suppressor-DNA Complex: Understanding Tumorgenic Mutations; *Science* (**1994**) *265*; 346-355.

141. J.T. Bolin, D.J. Filman, D.A. Matthews, R.C. Hamlin und J. Kraut; Crystal Structures of Escherichia Coli and Lactobacillus Casei Dihydrofolate Reductase Refined at 1.7 Angstroms Resolution. I. General Features and Binding of Methotrexate; *J.Biol. Chem.* (**1982**) *257*; 13650-13662.

142. A. Lavie, I.R. Vetter, M. Konrad, R.S. Goody, J. Reinstein und I. Schlichting; Structure of Thymidylate Kinase Reveals the Cause Behind the Limiting Step in AZT Activation; *Nat.Struct.Biol.* (**1997**) *4*(8); 601-604.

143. M. Marquart, J. Walter, J. Deisenhofer, W. Bode und R. Huber; The Geometry of the Reactive Site and of the Peptide Groups in Trypsin, Trypsinogen and Its Complexes With Inhibitors; *Acta Crystallogr.,Sect.B* (**1983**) *39*; 480-490.

144. K. Kinzler und B. Vogelstein; Life (and Death) in a Malignant Tumor; *Nature* (**1996**) *379*; 19-20.

145. C. Caelles, A. Helmberg und M. Karin; P53-Dependent Apoptosis in the Absence of Transcriptional Activation of P53-Target Genes; *Nature* (**1994**) *370*; 220-223.

146. W. El-Diery, S. Kern, J. Pietenpol, K. Kinzler und B. Vogelstein; Definition of a Consensus Binding Site for P53; *Nat.Genet.* (**1992**) *1*; 45-49.

147. P. Hainaut und J. Milner; Redox Modulation of P53 Conformation and Sequence-Specific DNA Binding *in Vitro*; *Cancer Res.* (**1993**) *53*; 4469-4473.

148. M. Hollstein, G. Moeckel, M. Hergenhahn, B. Spiegelhalder, M. Keil, G. Werle-Schneider, H. Bartsch und J. Brickmann; On the Origins of Tumor Mutations in Cancer Genes: Insights From the P53 Gene; *Mutat.Res.* (**1998**) *405*(2); 145-154.

9 Anhang

9.1 Schreibweise der Aminosäuren

Tabelle 9.1: Abkürzungsverzeichnis der 20 proteinogenen Aminosäuren.

Name der Aminosäure	Drei-Buchstaben-Code	Ein-Buchstaben-Code
Arginin	ARG	R
Lysin	LYS	K
Glutaminsäure	GLU	E
Asparaginsäure	ASP	D
Glutamin	GLN	Q
Asparagin	ASN	N
Serin	SER	S
Threonin	THR	T
Histidin	HIS	H
Prolin	PRO	P
Glycin	GLY	G
Alanin	ALA	A
Valin	VAL	V
Leucin	LEU	L
Isoleucin	ILE	I
Cystein	CYS	C
Methionin	MET	M
Phenylalanin	PHE	F
Tyrosin	TYR	Y
Tryptophan	TRP	W

9.2 Zusätzliche Ergebnistabellen und -diagramme

9.2.1 Größe, Oberfläche und Volumen der Proteinkomplexe

Tabelle 9.2: Größenangaben (jeweils Mittelwert und Standardabweichung) der Proteinkomplexe in den vier verwendeten Datensätzen.

	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Atome pro Komplex	5000 ± 4189	5782 ± 3999	7275 ± 3354	5755 ± 3301
Aminosäuren	322 ± 270	373 ± 259	478 ± 220	382 ± 211
Gesamtoberfläche [$Å^2$]	13652 ± 10895	15338 ± 10212	21277 ± 9789	15446 ± 8609
Bindungsoberfläche [$Å^2$]	1344 ± 2721	1326 ± 2858	2979 ± 1719	2092 ± 2510
Anteil an Gesamtoberfläche	6,3% ± 9,4%	5,5% ± 8,9%	13,5% ± 3,7%	12,8% ± 9,6%
Gesamtvolumen [$Å^3$]	44771 ± 37765	51945 ± 36164	64878 ± 29961	51586 ± 29645

9.2.2 Aminosäurenzusammensetzung von Bindungsbereichen

Tabelle 9.3: Mittelwerte und Standardabweichungen der Anteile der 20 proteinogenen Aminosäuren an der Oberfläche von Protein-Protein-Bindungsbereichen (Kapitel 5.4.3.1).

Aminosäure	Datensatz A	Datensatz B	Datensatz C	Datensatz D
ARG	8,4% ± 12,8%	9,1% ± 15,5%	6,6% ± 11,8%	8,6% ± 15,0%
LYS	6,2% ± 11,1%	7,5% ± 12,6%	5,5% ± 11,9%	4,8% ± 9,3%
GLU	5,8% ± 10,7%	6,4% ± 10,0%	4,2% ± 10,0%	3,7% ± 8,0%
ASP	4,6% ± 9,0%	5,1% ± 8,0%	4,9% ± 9,9%	2,8% ± 4,8%
GLN	4,7% ± 8,5%	3,5% ± 7,6%	4,9% ± 5,5%	4,5% ± 8,9%
ASN	5,0% ± 10,7%	4,6% ± 7,5%	4,6% ± 9,3%	3,9% ± 10,6%
SER	4,8% ± 8,5%	4,9% ± 7,4%	7,8% ± 9,9%	7,7% ± 10,3%
THR	5,5% ± 9,3%	5,4% ± 7,4%	5,2% ± 6,1%	4,8% ± 6,3%
HIS	3,3% ± 8,2%	3,2% ± 8,5%	2,9% ± 5,8%	3,5% ± 6,2%
PRO	4,6% ± 7,7%	5,1% ± 8,8%	5,3% ± 6,4%	5,4% ± 8,4%
GLY	3,7% ± 6,1%	3,7% ± 5,0%	4,4% ± 8,2%	6,8% ± 7,0%
ALA	4,3% ± 7,4%	3,7% ± 4,9%	2,3% ± 6,1%	4,8% ± 8,3%
VAL	5,8% ± 8,7%	5,3% ± 8,4%	3,6% ± 6,2%	4,6% ± 8,6%
LEU	9,5% ± 12,9%	9,0% ± 12,6%	7,2% ± 7,9%	7,6% ± 11,0%
ILE	4,9% ± 8,7%	5,4% ± 7,3%	2,1% ± 6,0%	3,8% ± 5,9%
CYS	1,7% ± 5,3%	1,8% ± 8,0%	0,6% ± 1,5%	4,2% ± 5,1%
MET	2,8% ± 7,2%	3,0% ± 5,9%	1,5% ± 3,8%	2,3% ± 6,1%
PHE	5,1% ± 8,0%	4,9% ± 8,3%	8,3% ± 9,3%	5,5% ± 9,9%
TYR	6,3% ± 10,0%	6,3% ± 9,4%	11,8% ± 12,2%	6,3% ± 9,1%
TRP	2,9% ± 7,8%	2,1% ± 6,7%	6,4% ± 7,9%	4,5% ± 9,0%

Tabelle 9.4: Gemittelte Anteile der 20 proteinogenen Aminosäuren an der Oberfläche von Protein-Ligand-Bindungsbereichen (Kapitel 5.4.3.2).

Aminosäure	Datensatz A	Datensatz B	Datensatz C	Datensatz D
ARG	8,4% ± 17,5%	8,8% ± 16,3%	8,3% ± 20,7%	11,8% ± 26,5%
LYS	4,8% ± 13,0%	6,3% ± 15,3%	0,9% ± 3,5%	5,0% ± 6,8%
GLU	4,1% ± 11,3%	6,1% ± 12,0%	0,7% ± 3,6%	6,7% ± 5,1%
ASP	5,9% ± 14,3%	5,6% ± 12,2%	3,1% ± 10,2%	2,2% ± 5,3%
GLN	3,3% ± 10,0%	2,8% ± 7,0%	0,8% ± 3,4%	0,1% ± 0,6%
ASN	4,7% ± 11,1%	5,9% ± 9,6%	5,0% ± 11,4%	3,5% ± 4,0%
SER	3,9% ± 9,5%	3,6% ± 6,8%	4,6% ± 11,3%	4,0% ± 5,1%
THR	3,7% ± 9,6%	3,9% ± 6,8%	3,6% ± 10,7%	2,6% ± 4,3%
HIS	4,5% ± 11,3%	3,9% ± 8,0%	9,2% ± 15,7%	2,4% ± 5,1%
PRO	2,0% ± 6,8%	2,6% ± 5,6%	3,0% ± 10,6%	0,7% ± 2,5%
GLY	5,7% ± 11,3%	5,0% ± 9,2%	4,4% ± 8,9%	11,7% ± 15,0%
ALA	5,5% ± 12,6%	4,4% ± 7,9%	1,1% ± 3,3%	6,9% ± 15,2%
VAL	5,0% ± 10,3%	5,2% ± 8,8%	5,1% ± 12,6%	10,9% ± 18,1%
LEU	7,5% ± 14,7%	7,0% ± 13,1%	4,5% ± 13,1%	9,7% ± 8,5%
ILE	6,2% ± 12,6%	5,0% ± 9,1%	0,4% ± 2,0%	3,2% ± 4,9%
CYS	1,4% ± 6,7%	1,4% ± 8,4%	0,2% ± 1,8%	0,6% ± 3,4%
MET	1,8% ± 6,6%	2,2% ± 5,6%	0,1% ± 1,0%	1,4% ± 2,6%
PHE	6,1% ± 13,7%	6,2% ± 12,6%	6,1% ± 13,7%	7,4% ± 9,9%
TYR	8,4% ± 15,3%	7,2% ± 11,5%	22,6% ± 25,8%	6,4% ± 13,7%
TRP	6,1% ± 14,9%	6,0% ± 12,7%	15,4% ± 24,9%	1,7% ± 6,3%

Tabelle 9.5: Gemittelte Anteile der 20 proteinogenen Aminosäuren an der Oberfläche von Protein-DNA-Bindungsbereichen (Kapitel 5.4.3.3).

Aminosäure	Datensatz A	Datensatz B
ARG	19,9% ± 17,5%	19,6% ± 14,8%
LYS	13,9% ± 10,7%	8,7% ± 9,1%
GLU	3,1% ± 7,2%	2,9% ± 4,7%
ASP	2,0% ± 4,1%	3,5% ± 6,9%
GLN	5,3% ± 9,4%	8,7% ± 13,6%
ASN	6,9% ± 7,5%	3,5% ± 5,5%
SER	8,6% ± 8,1%	7,6% ± 6,0%
THR	7,5% ± 8,0%	11,3% ± 8,1%
HIS	3,7% ± 6,7%	2,7% ± 5,2%
PRO	1,7% ± 5,6%	1,3% ± 3,9%
GLY	5,7% ± 6,9%	5,2% ± 6,8%
ALA	3,5% ± 5,4%	9,1% ± 11,0%
VAL	2,0% ± 3,9%	2,0% ± 4,0%
LEU	2,4% ± 4,7%	1,3% ± 3,5%
ILE	2,5% ± 4,7%	4,5% ± 8,2%
CYS	0,5% ± 1,6%	1,0% ± 2,2%
MET	2,2% ± 4,8%	0,4% ± 1,6%
PHE	3,0% ± 5,6%	3,4% ± 7,7%
TYR	4,3% ± 6,7%	2,6% ± 5,5%
TRP	0,9% ± 4,0%	0,4% ± 1,9%

9.2.3 Molekulare Oberfläche einzelner Aminosäuren

Tabelle 9.6: Mittlere Oberfläche (Mittelwerte und Standardabweichung) der Aminosäuren an der molekularen Oberfläche des Tripeptids Glycin-X-Glycin und den verschiedenen Bereichen der Proteinoberfläche (Kapitel 5.5).

Aminosäure	GLY-X-GLY [Å^2]	Gesamte Oberfläche [Å^2]	Protein-Protein-Bindungsbereich [Å^2]	Protein-DNA-Bindungsbereich [Å^2]	Protein-Ligand-Bindungsbereich [Å^2]
ARG	168,8	80,0 ± 36,2	96,3 ± 33,0	90,4 ± 37,2	49,3 ± 33,1
LYS	150,6	84,1 ± 28,0	90,8 ± 28,4	89,3 ± 28,0	48,3 ± 29,9
GLU	120,1	62,5 ± 26,2	69,2 ± 23,5	53,8 ± 24,5	32,3 ± 24,9
ASP	98,8	48,8 ± 23,8	56,4 ± 20,4	39,7 ± 23,5	30,7 ± 20,6
GLN	126,4	60,4 ± 29,2	70,3 ± 27,4	62,7 ± 26,1	41,6 ± 28,3
ASN	104,1	50,4 ± 26,4	59,6 ± 24,6	58,3 ± 21,1	38,7 ± 24,0
SER	82,0	36,7 ± 22,2	42,9 ± 18,8	43,2 ± 14,6	25,3 ± 16,7
THR	99,5	40,6 ± 25,2	53,1 ± 21,7	44,3 ± 17,1	29,3 ± 20,3
HIS	123,4	49,5 ± 32,0	64,4 ± 28,0	51,3 ± 24,3	31,7 ± 22,1
PRO	95,9	46,8 ± 25,8	57,1 ± 22,0	46,2 ± 24,7	32,4 ± 24,3
GLY	56,0	23,7 ± 16,9	31,5 ± 14,1	27,9 ± 10,1	23,5 ± 14,3
ALA	74,3	25,2 ± 21,6	38,7 ± 17,8	33,5 ± 15,5	24,7 ± 16,8
VAL	107,1	25,3 ± 26,0	48,1 ± 24,5	37,9 ± 21,1	28,6 ± 20,0
LEU	124,5	29,1 ± 30,0	59,4 ± 29,6	50,0 ± 33,1	32,8 ± 22,7
ILE	124,2	25,7 ± 28,3	55,2 ± 29,8	34,4 ± 28,6	31,8 ± 23,5
CYS	95,4	18,8 ± 21,2	42,4 ± 25,9	36,3 ± 22,9	25,0 ± 18,8
MET	128,5	35,2 ± 35,3	63,6 ± 32,2	50,5 ± 25,3	31,7 ± 22,8
PHE	140,2	33,5 ± 33,8	68,9 ± 33,7	50,5 ± 30,7	38,2 ± 25,3
TYR	149,7	45,6 ± 34,6	72,1 ± 33,9	60,1 ± 32,7	42,8 ± 28,6
TRP	167,7	44,7 ± 38,6	80,5 ± 38,6	61,5 ± 40,1	49,3 ± 28,1

Tabelle 9.7: Mittlere Oberfläche (Mittelwerte und Standardabweichung) der Aminosäurenseitenketten an der molekularen Oberfläche des Tripeptids Glycin-X-Glycin und den verschiedenen Bereichen der Proteinoberfläche (Kapitel 5.5).

Aminosäure	GLY-X-GLY [Å^2]	Gesamte Oberfläche [Å^2]	Protein-Protein-Bindungsbereich [Å^2]	Protein-DNA-Bindungsbereich [Å^2]	Protein-Ligand-Bindungsbereich [Å^2]
ARG	128,6	69,6 ± 30,5	83,7 ± 26,3	79,4 ± 29,9	43,7 ± 28,6
LYS	111,6	71,7 ± 22,1	77,0 ± 21,6	76,7 ± 20,7	42,7 ± 25,5
GLU	80,7	50,2 ± 20,1	55,4 ± 16,9	44,3 ± 18,9	26,2 ± 18,8
ASP	59,0	36,2 ± 17,1	41,6 ± 14,0	31,5 ± 16,1	23,9 ± 14,7
GLN	87,1	49,4 ± 23,3	57,8 ± 20,8	52,2 ± 19,7	35,1 ± 23,2
ASN	64,6	38,5 ± 19,5	44,6 ± 16,9	45,8 ± 13,6	30,4 ± 17,9
SER	42,7	23,4 ± 14,2	27,7 ± 11,5	30,5 ± 9,4	16,8 ± 11,3
THR	61,7	30,5 ± 19,1	38,8 ± 15,6	35,0 ± 13,8	22,2 ± 15,7
HIS	84,0	40,0 ± 25,8	52,5 ± 21,8	43,6 ± 20,8	27,3 ± 18,4
PRO	63,8	35,4 ± 20,3	43,1 ± 17,1	34,8 ± 21,0	23,7 ± 18,3
GLY	-	-	-	-	-
ALA	32,8	13,6 ± 12,2	21,6 ± 9,7	19,8 ± 9,8	14,1 ± 9,8
VAL	70,4	18,6 ± 20,4	37,2 ± 19,2	29,3 ± 18,6	23,0 ± 16,0
LEU	88,8	22,3 ± 24,4	48,1 ± 24,2	41,4 ± 28,9	27,2 ± 19,6
ILE	87,6	20,0 ± 23,3	44,5 ± 24,7	26,4 ± 26,3	26,2 ± 19,3
CYS	56,2	10,9 ± 14,3	27,7 ± 18,2	23,5 ± 18,1	16,6 ± 14,3
MET	90,9	27,1 ± 27,8	50,8 ± 25,2	44,2 ± 19,9	26,4 ± 19,1
PHE	101,5	26,4 ± 28,3	56,6 ± 27,4	41,7 ± 26,2	32,5 ± 22,4
TYR	110,3	38,5 ± 29,6	61,4 ± 28,3	51,9 ± 28,9	37,6 ± 24,7
TRP	130,4	37,6 ± 33,9	69,2 ± 33,9	51,5 ± 35,0	43,6 ± 25,6

9.2.4 Kontaktwechselwirkungen in Protein-Protein-Komplexen

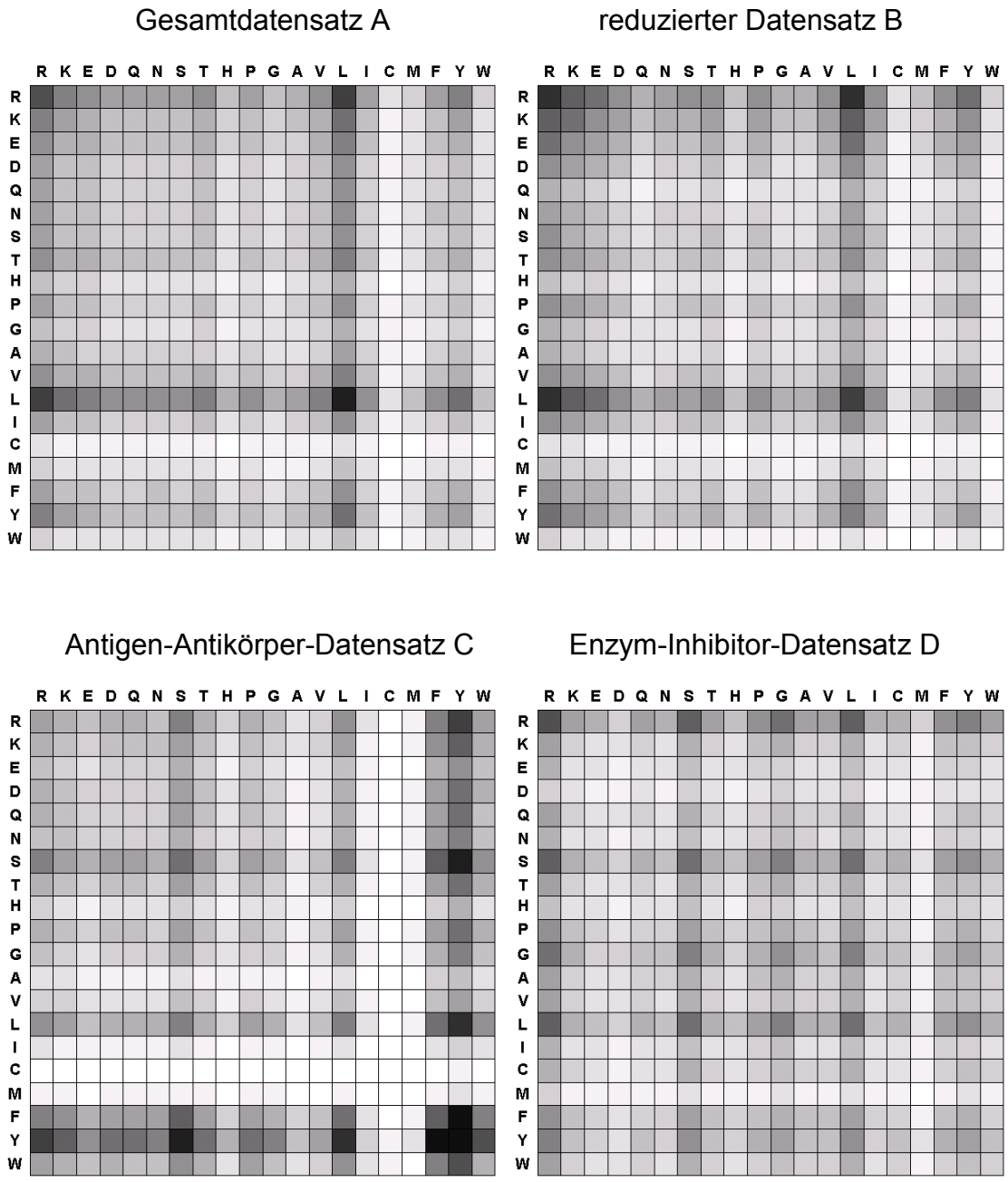


Diagramm 9.1: Theoretische Kontaktmatrix (Kapitel 5.6) der Aminosäurenwechselwirkungen zwischen Proteinen in den Datensätzen A-D (weiß: 0,0%; schwarz: 1,0%).

9.2.5 Untersuchungsergebnisse der molekularen Proteinoberflächen

Tabelle 9.8: Mittelwerte und Standardabweichungen der molekularen Eigenschaften der Oberflächendomänen (Teiloberflächen) an der molekularen Oberfläche (Gesamtoberfläche) von Proteinen (Kapitel 5.7).

Eigenschaft	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Anzahl der Domänen	1255853	75429	56646	14670
Größe der Domänen [$Å^2$]	261,7 ± 54,1	261,0 ± 53,1	260,6 ± 58,1	262,4 ± 53,6
Elektrostat. Pot. [kcal/mol·e]	-1,9 ± 34,8	-3,6 ± 34,4	1,4 ± 27,6	2,3 ± 33,5
Lokale Lipophilie	-0,0633 ± 0,0383	-0,0627 ± 0,0385	-0,0642 ± 0,0386	-0,0624 ± 0,0358
Flexibilität [$Å^2$]	26,0 ± 16,8	26,9 ± 14,9	26,9 ± 15,5	29,2 ± 14,7
H-Akz.-/H-Donordichte [$Å^{-2}$]	0,0440 ± 0,0115	0,0441 ± 0,0116	0,0461 ± 0,0110	0,0448 ± 0,0115
H-Akzeptordichte [$Å^{-2}$]	0,0229 ± 0,0064	0,0230 ± 0,0064	0,0235 ± 0,0060	0,0231 ± 0,0066
H-Donatordichte [$Å^{-2}$]	0,0212 ± 0,0086	0,0211 ± 0,0086	0,0226 ± 0,0076	0,0218 ± 0,0084
Tiefeninformation [Å]	1,30 ± 1,25	1,33 ± 1,25	1,02 ± 0,67	1,17 ± 1,04
Oberflächenkrümmung	2,28 ± 0,33	2,27 ± 0,33	2,35 ± 0,26	2,32 ± 0,33

Tabelle 9.9: Mittelwerte und Standardabweichungen der molekularen Eigenschaften der Oberflächendomänen in den Protein-Protein-Bindungsregionen der molekularen Oberfläche von Proteinen (Kapitel 5.7).

Eigenschaft	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Anzahl der Domänen	147784	7741	9442	2459
Größe der Domänen [$Å^2$]	264,7 ± 50,6	263,7 ± 49,9	253,5 ± 52,7	266,8 ± 50,0
Elektrostat. Pot. [kcal/mol·e]	-0,7 ± 32,1	-1,8 ± 32,3	-1,2 ± 28,4	-0,4 ± 33,4
Lokale Lipophilie	-0,0424 ± 0,0436	-0,0415 ± 0,0444	-0,0285 ± 0,0463	-0,0451 ± 0,0398
Flexibilität [$Å^2$]	21,0 ± 14,2	21,4 ± 11,2	21,4 ± 12,5	26,1 ± 13,4
H-Akz.-/H-Donordichte [$Å^{-2}$]	0,0381 ± 0,0116	0,0378 ± 0,0115	0,0385 ± 0,0123	0,0388 ± 0,0114
H-Akzeptordichte [$Å^{-2}$]	0,0199 ± 0,0062	0,0200 ± 0,0062	0,0203 ± 0,0072	0,0202 ± 0,0067
H-Donatordichte [$Å^{-2}$]	0,0182 ± 0,0084	0,0179 ± 0,0085	0,0183 ± 0,0076	0,0187 ± 0,0079
Tiefeninformation [Å]	1,24 ± 0,94	1,31 ± 0,97	1,02 ± 0,76	1,41 ± 1,16
Oberflächenkrümmung	2,33 ± 0,31	2,32 ± 0,31	2,33 ± 0,30	2,36 ± 0,38

Tabelle 9.10: Mittelwerte und Standardabweichungen der molekularen Eigenschaften der Oberflächendomänen in den Protein-DNA-Bindungsregionen der molekularen Oberfläche von Proteinen (Kapitel 5.7).

Eigenschaft	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Anzahl der Domänen	3273	244	3	0
Größe der Domänen [$Å^2$]	266,5 ± 56,4	263,5 ± 55,1	249,9 ± 74,7	-
Elektrostat. Pot. [kcal/mol·e]	50,3 ± 33,1	51,0 ± 29,4	17,0 ± 18,8	-
Lokale Lipophilie	-0,0626 ± 0,0339	-0,0608 ± 0,0364	-0,0370 ± 0,0538	-
Flexibilität [$Å^2$]	25,0 ± 15,5	25,3 ± 12,5	14,8 ± 0,6	-
H-Akz.-/H-Donordichte [$Å^{-2}$]	0,0500 ± 0,0121	0,0491 ± 0,0132	0,0457 ± 0,0149	-
H-Akzeptordichte [$Å^{-2}$]	0,0188 ± 0,0067	0,0181 ± 0,0071	0,0213 ± 0,0109	-
H-Donatordichte [$Å^{-2}$]	0,0312 ± 0,0090	0,0309 ± 0,0097	0,0240 ± 0,0071	-
Tiefeninformation [Å]	1,76 ± 1,48	1,51 ± 1,17	1,15 ± 0,39	-
Oberflächenkrümmung	2,23 ± 0,32	2,25 ± 0,33	2,33 ± 0,22	-

Tabelle 9.11: Mittelwerte und Standardabweichungen der molekularen Eigenschaften der Oberflächendomänen in den Protein-Ligand-Bindungsregionen der molekularen Oberfläche von Proteinen (Kapitel 5.7).

Eigenschaft	Datensatz A	Datensatz B	Datensatz C	Datensatz D
Anzahl der Domänen	10192	588	218	102
Größe der Domänen [$Å^2$]	250,0 ± 61,9	253,5 ± 60,0	246,1 ± 50,9	254,3 ± 55,7
Elektrostat. Pot. [kcal/mol·e]	9,0 ± 42,1	10,7 ± 43,7	-0,9 ± 34,4	9,4 ± 56,3
Lokale Lipophilie	-0,0264 ± 0,0488	-0,0265 ± 0,0505	-0,0147 ± 0,0467	-0,0458 ± 0,0518
Flexibilität [$Å^2$]	19,1 ± 11,4	19,4 ± 10,7	16,7 ± 7,9	19,7 ± 8,2
H-Akz.-/H-Donordichte [$Å^{-2}$]	0,0495 ± 0,0184	0,0492 ± 0,0206	0,0415 ± 0,0115	0,0435 ± 0,0135
H-Akzeptordichte [$Å^{-2}$]	0,0244 ± 0,0095	0,0245 ± 0,0100	0,0211 ± 0,0073	0,0234 ± 0,0085
H-Donatordichte [$Å^{-2}$]	0,0251 ± 0,0131	0,0248 ± 0,0139	0,0204 ± 0,0077	0,0201 ± 0,0100
Tiefeninformation [Å]	5,07 ± 3,21	4,68 ± 3,15	1,85 ± 1,41	4,52 ± 3,24
Oberflächenkrümmung	1,63 ± 0,47	1,68 ± 0,51	2,06 ± 0,35	1,78 ± 0,63

9.2.6 Trainingsergebnisse des neuronalen Netzes I

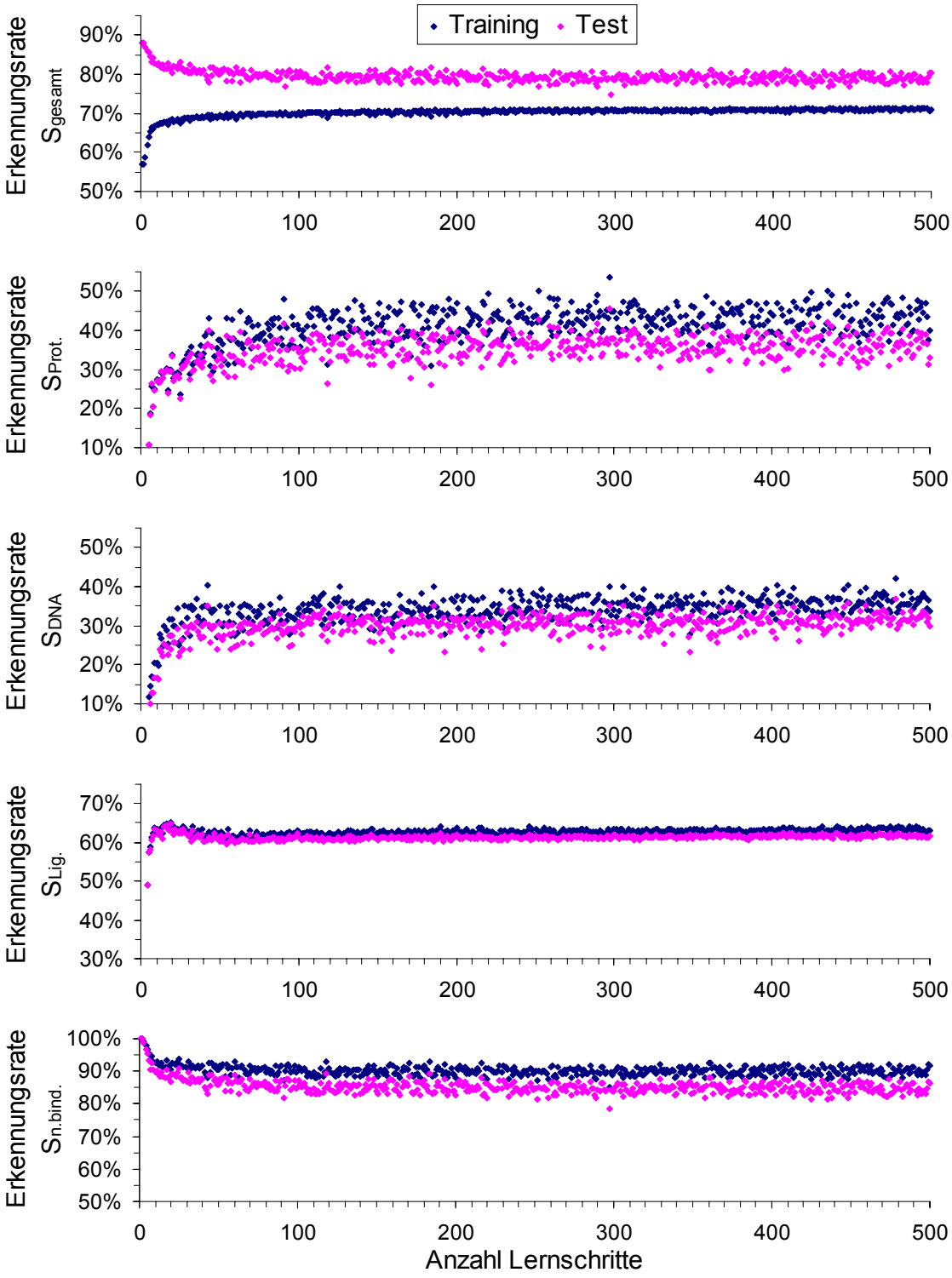


Diagramm 9.2: Lernverhalten des neuronalen Netzes I im Trainings- und Testdatensatz (Kapitel 6.2.4): Dargestellt sind von oben nach unten die Gesamterkennungsrate S_{gesamt} und die Erkennungsraten der vier Oberflächenklassen (Protein-, DNA-, Ligandbindungsbereich und nichtbindende Oberfläche).

9.3 Hilfsmittel

- Der vorliegende Text wurde mit dem Textverarbeitungsprogramm Microsoft Word 2000 geschrieben.
- Teile der Auswertung sowie die Erstellung aller Diagramme wurden mit den Programmen Excel 2000 von Microsoft und Origin 5.0 der Firma Microcal vorgenommen.
- Zur Anfertigung der schematischen Darstellungen diente sowohl Paintshop Pro 6 von der Firma Jasc Software als auch das Programm Micrografix Designer 6.0.
- Die Strukturen der Tripeptide wurden mit dem *Molecular Modeling* Programmpaket SYBYL 6.7 [125] von Tripos erzeugt.
- Bei der Bearbeitung der Proteinstrukturen wurde das Programm CHARMM 24 [95] verwendet.
- Die Entwicklung des neuronalen Netzes erfolgte unter Zuhilfenahme des Stuttgarter Neuronale Netze Simulators SNNS Version 4.2 [135].
- Die Molekülgraphiken wurden mit MOLCAD II erstellt. MOLCAD II wurde von mir zur Visualisierung und Analyse molekularer Eigenschaften auf der Basis des im Arbeitskreis entwickelten MOLCAD [86] neu implementiert und weiterentwickelt.

- Für die Berechnung und Analyse der Daten standen folgende Computer zur Verfügung:
  - Silicon Graphics PowerOnyx R10000 am Hochschulrechenzentrum der TUD
  - IBM RS/6000 H70 und J50 am Hochschulrechenzentrum der TUD
  - Silicon Graphics Indigo und Indy am Institut für Physikalische Chemie
  - Dell Precision 410 und Dell Poweredge 2400 (Intel Pentium III, SuSE Linux 7.0) am Institut für Physikalische Chemie

Lebenslauf

Name:	Matthias Keil
Geburtsdatum:	13. April 1971
Geburtsort:	Bensheim
Staatsangehörigkeit:	deutsch

Schulbildung: 1977 - 1981
Nibelungenschule Heppenheim

1981 - 1990
Starkenburg-Gymnasium Heppenheim
Abschluß: Abitur

Wehrdienst: 1990 - 1991
Grundwehrdienst

Studium: 1991 - 1996
Chemiestudium an der Technischen Hochschule Darmstadt
Diplomarbeit am Institut für Physikalische Chemie I in der Arbeitsgruppe von Prof. Dr. J. Brickmann
Thema: „Computergestützte Untersuchungen der Bindungsregion des p53-Protein-DNA-Komplexes“
Abschluß: Diplom-Ingenieur

1996 – 2002
Übernahme der Entwicklungsarbeit am *Molecular-Modeling*-Programm MOLCAD in der Arbeitsgruppe von Prof. Dr. J. Brickmann und Promotionsstudium an der Technischen Universität Darmstadt unter der Leitung von Prof. Dr. J. Brickmann
Thema: „Modellierung und Vorhersage von Strukturen biomolekularer Assoziate auf der Basis von statistischen Datenbankanalysen“

Matthias Keil
Marienbader Straße 6a
64646 Heppenheim

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides Statt, daß ich meine Dissertation selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe.

Darmstadt, den 12. Februar 2002

Matthias Keil

(Matthias Keil)